

Hollow apologies should be avoided

The head of the Max Planck Society is right to resist pressures to apologize for the actions of a previous generation. But acknowledgement and condemnation cannot come too soon.

“It is not within the moral authority of those who did not directly take part [in Nazi experiments on humans] to ask forgiveness for unforgivable crimes on behalf of others.” Writing in the weekly newspaper *Die Zeit*, Hubert Markl, president of the Max Planck Society (MPS), puts his finger on a quandary that continues to deeply trouble many in the German scientific community: how should the postwar generation of scientists deal with the crimes committed in the name of medical research during the Third Reich?

Most scientists seem to agree that formal apologies to surviving victims of the experiments is in order, and long overdue. In the first decades after the war a conspiracy of silence reigned over the role of scientists supported by the Kaiser Wilhelm Society (KWS, the forerunner of the MPS) and the Deutsche Forschungsgemeinschaft (DFG) in such experiments. Silence was important to the scientists who moved into key positions in the postwar scientific community. It was broken in the 1980s when academics such as geneticist Benno Muller-Hill published evidence linking scientists who continued to head research institutes after the war to these crimes.

Pressure on the MPS and the DFG to apologize to the survivors of such experiments while they are still alive has increased, particularly since the 55th anniversary of the freeing from Auschwitz a few weeks ago. Many academics argue that Willy Brandt was prepared to kneel at the Warsaw ghetto 30 years ago and ask forgiveness for the crimes against Jews committed by Germans during the Third Reich (indeed, German president Johannes Rau only last week apologized in the Israeli parliament): surely representatives of research organizations should be prepared to do the same?

Apology too easy

However, what is generally accepted as appropriate action for Brandt, elected chancellor of a nation whose majority had lived through the war, appears not to be a simple matter to the war babies who now run German research organizations. Markl argues that apology is too easy and gives more to those who apologize — in particular an easing of conscience — than to those asked to accept the apology.

Markl's views have not brought him popularity in a society that struggles with the concept of collective guilt. But his reasons for holding back are based on a morality that has been well thought through, resists emotional pressure and detachedly considers the position of those who are not personally responsible and of those, still alive, who suffered.

The crimes committed are too great to allow forgiveness, he argues. Baby-boomers should not trivialize them by seeking forgiveness on behalf of others now dead, who, for all we know, might not even regret what they did. Apologies are not useful to the victims. More important is an explicit portrayal of what happened. At Markl's behest, the MPS recently launched a five-year research programme, run by science

historians who are completely independent of the MPS, into the history of the KWS during the Third Reich. Markl, who says that this should have happened years ago, wants to wait for the results of the programme before deciding if an apology is due, or helpful.

Markl is alone among the heads of Germany's main research organizations in his stand against apology. His position angers some who see a moral duty to apologize, and bemuses others who know that it may be easy, and even self-serving, to apologize, but feel that it cannot be intrinsically wrong, and at least nothing would be lost. But his arguments are generally sound.

Consult survivors

Nevertheless, the scientific community's appropriate sensitivity to the crimes of its predecessors cannot appear to be ignored. That could lead to accusations that the MPS is in effect perpetuating a conspiracy that protected KWS scientists who continued their research careers with the MPS after the war — exactly what Markl wanted to avoid by making the research group independent, and well-resourced.

Moreover, Markl appears to confuse his own arguments by not separating the issue of apology from the work of this independent research group. There is already sufficient evidence for the shameful involvement of KWS scientists in Nazi experiments, and even 50 years of research will not provide a comprehensive and definitive picture of what happened.

Ernst-Ludwig Winnacker, president of the DFG, has suggested a different way around the conundrum. The DFG is continuing to investigate its own role in funding research that supported Nazi policies. Winnacker believes that there is already enough evidence to warrant an apology from the DFG before investigations are complete. But, like Markl, he also wonders if an apology would be helpful. His solution is to contact survivors and ask them if they want an apology. After all, he says, survivors should not be used as victims again, to return to grace organizations whose fate may not interest them.

A mistake that both Markl and Winnacker — indeed, the whole community — could be making is to focus the debate so strongly on apology. Apology misguidedly associates the postwar generation directly with the activities of its predecessor. An expression of a sense of shame would be less of a hollow gesture, as it is a meaningful act to acknowledge and explicitly react to a horrific aspect of one's scientific heritage. The real issue, however, is that of their public recognition and condemnation of actions by scientists against Jews and other outcast groups and populations. Apologies may be due, rather, for the MPS's and DFG's ignoring of the issue until quite recently.

Were Markl and Winnacker publicly to issue to survivors statements of recognition and condemnation, they would demonstrate the total lack of acceptance by today's German scientific community of exploitation of humans in scientific experiments. It might also grant peace of mind to those in the scientific community haunted by the acts of a previous generation. ■





Drosophila
Shotgun approach
sequences the
fruitfly genome
p817



Commerce
Harvard Med School
may relax rules
on business links
p818



GM maize
German government
blocks cultivation
of Bt maize
p821



Heretic?
Supporters take to
streets in memory
of Giordano Bruno
p822

Celera in talks to launch private sector human proteome project

Paris

Celera Genomics, the private-sector rival to the publicly funded international effort to sequence the human genome, is now said to be contemplating an industrial-scale Human Proteome Project, this time well ahead of any comparable public venture. The goal would be to identify the properties and functions of every protein expressed in an organism: its 'proteome'.

Celera is in advanced discussions with Denis Hochstrasser, one of the founders of GeneBio — the commercial arm of the Swiss Institute of Bioinformatics, which is among the world's leading protein research centres — and an authority on two-dimensional polyacrylamide gel electrophoresis (2D PAGE) technology. The goal is to combine new technologies with the workhorse 2-D electrophoretic gel technique to develop an industrial-scale attack on the proteome.

Reaching a deal now seems to hang on the question of data access. Amos Bairoch, a co-founder of GeneBio, says that joining up with a private company will result in restrictions on data being placed in the public domain. He would prefer the Swiss groups to collaborate with a publicly funded international human proteome project, with all data being put rapidly into the public domain.

As the sequencing of the human genome nears completion, characterizing the human 'proteome' is the next big task (see *Nature* 402, 715–720; 1999). One approach is to try to predict proteins directly from gene sequences. In practice, however, this is riddled with problems.

Post-translational modifications, for example, mean a single gene sequence can yield over 20 structurally different proteins. So researchers instead want to identify the actual proteins found in any biological sample.

The many difficulties in scaling up proteomics, however, have made an industrial-scale approach impractical up until now. Protein spots are often still excised by hand from gels, for example, for further analysis, while hydrophobic or big proteins do not dissolve in the solvents used in 2-D PAGE.

That may now be about to change. Last November, Hochstrasser and colleagues at

the Swiss Institute of Bioinformatics published a blueprint for a 'molecular scanner' that simultaneously digests proteolytically the thousands of proteins on a gel, and electro-transfers them onto a membrane, in a procedure akin to Southern blotting (see *Anal. Chem.* 71, 21; 1999).

The peptides on the membranes are then measured by laser mass spectrometry, and the protein identified by matching this 'peptide mass fingerprint' automatically over the Internet with entries in a protein sequence database. The result is an annotated digital map of the proteins on a gel — a snapshot of any sample — that is available online. This database can be searched by protein name, or by peptide fingerprint.

Hochstrasser says Perkin-Elmer Corporation, which produces mass spectrometers, has agreed with GeneBio to develop the 'molecular scanner'. Significantly, Perkin-Elmer is also about to release a next-generation series of mass spectrometers, touted as being a major improvement on existing machines.

Meanwhile, Craig Venter, president of Celera Genomics — a subsidiary of Perkin-Elmer — has been wooing Hochstrasser with a view to a joint venture. This may combine hundreds of the new machines with the Swiss technology in a factory approach that would seem to amount to nothing less than an all-out human proteome project.

"Nothing has been signed yet," says Hochstrasser, adding that GeneBio, which is committed to public access, would not be involved. A "new entity" would channel the Swiss groups' expertise, he says. But, he adds, "There is no question [that a deal with Celera] would lead to a major scale-up."

He reckons that the problem of hydrophobic proteins can also be overcome by first precipitating them and redissolving them in alternative solvents. But, he says, the complexity of proteomics means that, although an industrial-scale approach could yield enormous amounts of data, alternative technologies will also be needed to finish the job.

Bairoch would prefer the Swiss institute

Eros begins to yield up its secrets

Washington

After making its Valentine's Day date with the asteroid Eros last week, the US space agency NASA's Near Earth Asteroid Rendezvous (NEAR) spacecraft has begun its year-long study of the tumbling space rock.

Some early results have confirmed ground-based studies and a run past Eros in December 1998. The asteroid is solid, rather than a conglomeration of rubble. And the first spectral scans of its mineral content — much better ones are to come — confirm that one side of Eros is higher in pyroxene than the other.

But there have been surprises, too. The surface is more heavily cratered — and therefore older — than expected. And scientists on the mission are puzzling over bright regions on the surface that are probably different varieties of rock.



Head over heels: Eros tumbles through space.

The first pictures also show ridges within craters that may reflect geologic layering from Eros' parent body — if, indeed, the asteroid is a fragment of a larger object.

The spacecraft will lower itself into a circular orbit 100 km from the centre of the asteroid in mid-April, and will then drop to a 50-km orbit in May.

Tony Reichhardt

to work with a publicly funded international consortium along the lines of the international human genome project. "It is a big debate in Geneva," he says. But he admits that this may be difficult. Whereas the US National Institutes of Health is playing the largest role in the human genome effort, Europe is strong in proteomics and would probably demand a leading role in any international effort.

At the same time, the political fragmentation of Europe means that there is little realistic hope of Europe getting organized quickly, says Bairoch. He points, for example, to the fiasco when the European Union almost let its foremost bioinformatics structure, the European Bioinformatics Institute, fall into bankruptcy (see *Nature* 402, 3–4; 1999). And when the publicly funded Swiss-Prot, one of the world's leading protein databases, faced closure for lack of money (see *Nature* 381, 266; 1996), it was only rescued by the creation of GeneBio, which ploughs 75 per cent of its profits back into the Swiss Institute of Bioinformatics.

A deal between Celera and Hochstrasser might prod the public sector into action, he adds. "I think the combination of having Celera in the human genome race was ultimately beneficial for all," says Robin Offord, a director of GeneBio.

Hochstrasser says, "We will try to make as much available for free as we can." But he argues that a compromise — such as an agreed delay in releasing data — may be needed to satisfy the demands of the private sector. "It is difficult to find the balance," he says. "We haven't signed yet because we would like free access to protein data. That is why it won't be GeneBio, but a new outfit."

Officials from the human genome project in the United States and Britain said comment would be premature since this was the first they had heard of a private-sector human proteome initiative.

Neither Celera nor Perkin-Elmer was available for comment. Heather Kowalski, corporate communications manager at Celera Genomics, explains that the companies are in a "quiet period" in the run-up to a secondary stock offering of 1.6 million Celera shares in early March.

The prospectus states that the offer is to fund "Celera's new product and technology development activities in functional genomics, with an emphasis on proteomics," as well as "general corporate purposes including possible acquisitions, alliances or collaborations". **Declan Butler**



Bairoch: fears loss of public access.

Biomedical centre memorial to victims of Nazi research

Munich

The Max Delbrück Centre for Molecular Medicine (MDC) in Berlin, Germany's largest biomedical research centre, is to erect a memorial to the victims of the Nazis' 'euthanasia' programme, using DM1.5 million (US\$750,000) from the country's national lottery.

The MDC has planned a memorial for several years. It occupies the campus that once housed the Kaiser Wilhelm Institute (KWI) for Brain Research. Before the Second World War, the institute was one of the world's leading neuroscience research facilities — only recently has its involvement in Nazi experiments become public knowledge.

Detlev Ganten, the director of the MDC, says that the memorial, which depicts a handicapped child, will bear an inscription referring to this history and reminding readers of the ethics of clinical research. "A patient becomes a research object, and this difficult situation can very easily become abused," he explains.

The KWI's director during the Third Reich, Hugo Spatz, and his younger scientific collaborator, Julius Hallervorden, did research using brains from victims of the 'Aktion T4' — one part of the Nazi programme to exterminate *lebensunwertes Leben* (life not deemed worthy of living).

After the war, a US army officer interrogated Hallervorden. He admitted that when he learnt of the exterminations, he asked for brains to be removed and preserved so that "good use could be made of the material".

But he was never charged with any crime. Both Hallervorden and Spatz continued their careers and became respected members of Germany's scientific community. Hallervorden died in 1965, and his activities during the Nazi era only resurfaced in 1983, when his collection of brain sections — including material from children killed in the Aktion T4 — was rediscovered in the cellars of the Max Planck Institute for Brain Research in Frankfurt (see *Nature* 339, 498; 1989).

Correspondence between personnel at the Dachau concentration camp and SS leader Heinrich Himmler shows that Spatz both knew about and participated in the use of victims of the Nazis in research. It also reveals that he used brains from prisoners killed in Dachau in his research.

In the light of this information, there was a move three years ago to rename Hallervorden-Spatz disease — a rare childhood hereditary neurodegenerative disorder — although this has not yet happened. But last year the name of the Hugo Spatz prize, awarded by the German Neurological Society, was discreetly changed to the Adolf Wallenberg prize, after the eminent Jewish-German neurologist who was forced to emigrate in 1938.

The plans for the memorial coincide with renewed efforts by Germany's science organizations to come to terms with their activities under the Nazis (see *Nature* 403, 474–475; 2000). A meeting on this theme will be held in May at the Bundesarchiv (federal archives) in Berlin.

Patrick Weydt

Australian industry 'starving' R&D

Sydney

Australian scientists are as adept at commercializing their academic work as their US counterparts, according to a national survey — but industry continues to reduce its commitment to research and development.

The results were released two weeks ago, just before Australia's first National Innovation Summit. In 1998, the members of the Australian Tertiary Institutions Commercial Companies Association, which represents the business arms of 48 universities, evaluated the commercial potential of 274 inventions, filed 161 patent applications and were granted 103 patents.

Australian universities launched 46 successful start-up companies between 1996 and 1998. This equates to one company for

every A\$100 million (US\$63 million) of university research funding — close to the average that a 1997 study found for US universities.

But Australia ranked 19 out of 24 leading economies in its business spending on research and development, as a percentage of gross domestic product (GDP), during 1996–97. The most recent estimate predicts that there was a further 10 per cent drop in such spending during 1998–99.

Contributing to government chief scientist Robin Batterham's review of the Australian science base, Vicki Sara, chair of the Australian Research Council, said that spending on university research has dropped by 13 per cent as a percentage of GDP over the four years of the coalition government.

Peter Pockley

Merck blocks 'safer' gene therapy trials

Washington

Gene therapy researchers at the Baylor College of Medicine in Houston, Texas, who recently reported success with the 'gutless' adenoviral vector in animals may not get to test the vector in humans.

The pharmaceutical company, Merck, which owns the licence on a technique used to produce the vector, is refusing to extend its material transfer agreement (MTA) with Baylor, saying that it does not want to be liable for any problems the vector might cause in clinical trials.

Conventional adenoviral vectors usually have only one or two small regions deleted, but 'gutless' (also known as 'fully deleted' or 'helper dependent') adenoviral vectors have a large chunk of their viral genome excised. This means that they could be used to avoid two problems that have limited the usefulness of other adenoviral vectors: inflammation and lack of long-term expression.

Viral genomes in adenoviral vectors trigger immune responses, including inflammation. This is thought to have played a role in the death of 18-year-old Jesse Gelsinger, the first death directly attributed to gene therapy (see *Nature* 401, 517; 1999). In a later hearing, Merck's former senior vice-president Thomas Caskey testified that the 'gutless' vectors might provide a safer alternative to

earlier generations of adenoviral vectors.

Animal experiments reported by Arthur Beaudet, chairman of molecular and human genetics at Baylor, at a conference in Colorado last month, reinforce that view, says Beaudet. In these experiments, Lawrence Chan, professor of molecular and cellular biology and medicine at Baylor, tested the vector in mouse models for inherited high cholesterol.

The experiments resulted in gene expression and lowered plasma cholesterol levels for at least six months. Whereas controls treated with a conventional adenoviral vector had marked inflammation in the liver, says Chan, mice that received the 'gutless' adenoviral vector experienced no significant inflammatory response.

After the experiments but before Gelsinger's death, Beaudet asked Merck to extend the MTA for a tool used to manufacture the 'gutless' vectors. The tool removes a large chunk of the vector's genome, yet still allows it to replicate. Merck denied the request, citing its concerns about liability.

The tool in question emerged from efforts in the early 1990s by a number of scientists — including Caskey, who was then at Baylor — to develop 'gutless' viruses that needed a second, 'helper' virus to replicate. A research group headed by Frank Graham,

professor of biology and pathology at McMaster University in Canada, subsequently solved the problem of eliminating the helper virus in the final vector preparations by excising the viral packaging signal.

Merck, which bought a licence on Graham's patent, retains intellectual property rights on it. Even if Baylor researchers sign a liability waiver, the US Food and Drug Administration could still hold the company to "ultimate regulatory responsibility", says Larry Hirsch, vice-president of public affairs at Merck Research Laboratories in Whitehouse Station, New Jersey.

Merck's refusal to extend the Baylor MTA for the vector-producing tool beyond pre-clinical use is "not unique to gene therapy", says Hirsch. "It's the way we handle all products in early stages of development." He adds that the vector is "nowhere near" ready for testing in a clinical trial.

Chan says the refusal is "frustrating". The Baylor team's only alternative, beyond abandoning the 'gutless' adenoviral vector, is finding a new way to create one. While that remains a possibility, it would probably require years of work. Producing large enough quantities of the vector for humans remains a problem, adds Graham. Meanwhile, Merck plans to continue developing and testing the vector on its own. **Paul Smaglik**

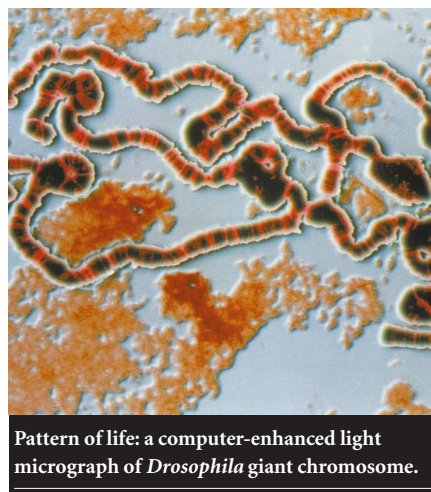
Celera's shotgun approach puts *Drosophila* in the bag

Washington

Celera Genomics this week announced that it has sequenced the genome of the fruitfly *Drosophila melanogaster*. The announcement was made at the annual meeting of the American Association for the Advancement of Science, and the sequence will be published next month in the journal *Science*.

Celera researchers claim that their success in sequencing such a large, complex eukaryotic genome is a vindication of their whole-genome — or shotgun — approach to sequencing (see *Nature* 401, 729; 1999). Gene Myers, head of Celera's computational biology group, says they met their goals simply on the basis of assembling the sequence data.

These goals included sequencing at least 95 per cent of the euchromatin (that part of the genome containing the vast majority of genes), meeting standards for accuracy and completeness accepted by the National Institutes of Health for human sequencing, and having no gap larger than 150,000 base pairs and an error rate of less than 1 in 10,000.



Pattern of life: a computer-enhanced light micrograph of *Drosophila* giant chromosome.

In the programme abstract for the meeting, Myers wrote that preliminary trials suggest that 99.97 per cent or more of the euchromatic portion of the genome will be assembled.

Celera collected more than 3 million shotgun paired reads from the *Drosophila* genome over a four-month period ending last September. Of the 180 million base-pair

(Mbp) genome, 120 Mbp was sequenceable DNA. Work began last November to close the 1,600 gaps in the sequence. These made up 2.65 Mbp of gaps straight out of the assembler. Their efforts may well "continue for another year or more," says Adams.

Celera say that most of these gaps are places where they have no data, rather than gaps caused by repeat sequences. It has been estimated that there are roughly 13,601 genes in the genome, although this number is expected to "evolve" over the months ahead.

Myers estimates that, from Celera's experience with *Drosophila*, the problem in obtaining the human sequence will be the larger scale of the genome rather than the assembly of sequence data.

In the fly, an assembler computes overlaps between reads in less than 13 hours on a four-processor platform. Myers says that by arranging assembly to be concurrent with data collection, Celera's assembler should achieve a comparable result in the human genome with a three-month computation, and on a ten-processor platform. **Natasha Loder**

Med school to relax rules on business links?

Boston

A report by senior faculty members at Harvard Medical School has recommended easing conflict-of-interest guidelines to give researchers greater flexibility in their commercial dealings and more opportunity to profit from their work.

The proposals are said to reflect concern that researchers are leaving for institutions with less stringent rules on stockholding in outside companies and earnings from consultancy fees.

Under the medical school's current guidelines, which were set in 1990 and modified in 1993, faculty members cannot own more than \$20,000 of stock in companies for which they do research. Furthermore, consulting fees from an individual company cannot exceed \$10,000 per year.

In addition, researchers cannot spend more than 20 per cent of their time working outside the university. Nor can they enlist students or postdoctoral fellows in their corporate activities.

The new guidelines are said to bring the monetary limits into line with those at other leading universities. "The current limits on equity ownership and consulting are quite low by contemporary standards," says David Blumenthal, a physician and health-policy specialist at the medical school and Massachusetts General Hospital. "It's appropriate that they be increased."

Even with the revamped policy, he notes, "Harvard will not be more permissive than other places like Stanford and Johns Hopkins". John Lacey, a spokesperson for the medical school, says that it has "some of the most stringent guidelines in the country", adding that "researchers are going to other institutions because they have more relaxed guidelines".

But Jonathan Beckwith, a microbiologist at the medical school, doubts this, saying it may just be an excuse to relax the guidelines. "This place is growing like mad. We're attracting stars from all over the world, and I haven't heard of any cases of people leaving."

Sheldon Krimsky, a health-policy specialist at nearby Tufts University, agrees. "Given how hard it is to get an appointment there, it's hard to believe that many people would give up a Harvard job for the freedom to pursue business ventures. It's even harder to believe that Harvard is worried about people leaving, because they have their pick."

The medical school faculty is split over the issue, according to one senior researcher who declines to be identified. "There's a real division of opinion here," he says. "Where you come down depends on what personal interests you have at stake. Those most in favour of the new policy want a bigger



Massachusetts General Hospital: tight guidelines may be causing the loss of top researchers.

financial return for the work they do for companies."

Krimsky, who attended a meeting at Harvard last month where the revised guidelines were discussed, argues that any steps to ease conflict-of-interest rules are worrisome "because we're already seeing some bad effects". He says that scientists are becoming more secretive, impeding exchange of information or cell lines, and that publication is

being delayed to protect proprietary concerns (see *Nature* 398, 359; 1999).

"There are already many problems, and the proposed changes would just exacerbate them," he says. "Harvard should set the highest standards for ethics in research, since they set the highest standards when it comes to offering positions to their faculty."

Hank Greely, a law professor at Stanford University and co-director of its Program in Genomics, Ethics and Society, says that it is healthy for institutions like Harvard to review their ethics policies regularly, but that a perfect solution is impossible. "As long as there are interactions between commercial sectors and academia, you have to worry about conflicts of interest. But breaking off these interactions would be bad for society."

Nevertheless, the unnamed researcher worries that the lure of money can divert universities from their primary mission — "getting the truth". This, he says, is especially true at an institution such as Harvard, whose motto is simply *Veritas*.

The report, drafted by a panel headed by Eugene Braunwald, professor of medicine and of ambulatory care and prevention at the medical school, is not being made public. But officials say that its recommendations will receive further consideration from the school's faculty and administrators before they are adopted.

Steve Nadis

Spanish boost for young scientists

Barcelona

The Spanish government, sensitive to criticism over the difficulties of providing attractive career opportunities for young researchers, has approved a programme to improve the pay and working conditions of university assistant professors.

The programme, which came into effect last month, removes a limitation preventing universities from employing full-time assistant professors for more than three years. They can now have their contracts renewed as many times as necessary if they pass an external evaluation.

In addition, new recruitment procedures have been introduced for full-time assistant professors with PhDs, who will receive the same salary — 4 million Ptas (US\$23,700) a year — as associate tenured professors.

The move is expected to benefit more than 10,000 assistant professors and to cost the government 21 billion Ptas over three years. Autonomous regional governments will have to contribute a further 7.7 billion Ptas over this period.

The government's actions have been

welcomed by labour unions representing university researchers. But left-wing parties and some university rectors are concerned at the financial implications and the lack of medium- and long-term provisions. Some observers argue that the award of the extra money may be politically motivated, as a general election is due to take place next month.

Antonio Caparrós, rector of the University of Barcelona, says that the government's action are likely to stabilize the position of assistant professors. But he cautions that problems involving university professors require long-term solutions.

Meanwhile, Jorge Fernández-Díaz, Spain's secretary for education, has promised to deliver a set of university reforms to parliament by the end of the year. These are expected to include a proposal to reduce from two to one the number of local university representatives on five-member appointment committees — a controversial aspect of Spain's recruitment procedures (see *Nature* 396, 712; 1998).

Xavier Bosch

AAAS members fret over links with theological foundation

Washington

Leading members of the American Association for the Advancement of Science (AAAS) are questioning the nature of its involvement in a programme on science and religion backed by the controversial John Templeton Foundation.

The foundation promotes “a path of cooperation” between the sciences and religion, funding up to 100 interdisciplinary courses each year and publishing a newsletter called *Progress in Theology*. Its other beneficiaries include an annual \$1.24 million prize for progress in religion, and a study of the benefits of private enterprise called the Freedom Project.

The AAAS's physics section forced a public review of the association's \$2 million Program of Dialogue on Science, Ethics and Religion at the annual meeting of its governing council last Sunday (20 February). Some questioned the programme's independence from the Templeton Foundation, and challenged its consistency with the association's aims.

But they accepted assurances from AAAS president Stephen Jay Gould and Al Teich, head of its science policy division, that steps are being taken to bolster the programme's independence.

The programme has supported a number of activities aimed at developing a dialogue between science and religion. It was set up four years ago with initial support from the John Templeton Foundation. So far, it has attracted \$1.1 million from the foundation and \$950,000 from other sources.

The physicists are claiming a conflict of interest between the AAAS, whose mission is to promote science, and the Templeton Foundation, which — said Hans Frauenfelder, a physicist at the Los Alamos national laboratory in New Mexico and past chairman of the AAAS physics section — believes there will ultimately be a “rapprochement” between science and religion.

“We are concerned that the Templeton Foundation has exercised undue influence on the programme,” adds John Peoples, former director of Fermilab in Illinois and current chairman of the AAAS physics section. “At a bare minimum, we feel that there is a conflict between the values and goals of the Templeton Foundation and those of the AAAS.” Frauenfelder presented data suggesting that speakers at meetings convened by the programme were often associated with the Templeton Foundation.

According to Rolf Sinclair, past secretary of the physics division, six people have served on the advisory boards to both the AAAS

programme and the foundation. Sinclair says that several nominations by the physics division to the advisory board have been summarily rejected. “I don't think it is an advisory committee: it is a committee of people involved in the programme,” he says.

But Al Teich, head of science policy at the AAAS, defended the programme vigorously. “This is a very religious country,” he told the council. “Religious leaders are listened to. Their views are very important and we need to talk to them if science is going to be influential and take a role in shaping our lives.”

Teich said that the programme was currently applying for \$500,000 of additional outside support. If obtained, this would allow it to reach its objective of restricting Templeton Foundation support to 40 per cent of the total.

Joel Primack, a physicist at the University of California at Santa Cruz and the new chairman of the programme's advisory committee, says that the number of its members who also advise the Templeton Foundation had been cut to one.

He cited the programme's interest in environmental matters, which Templeton refuses to support, as evidence of its independence. “The areas we think are important are in no way influenced by the Templeton Foundation,” he says.

But Gould came close to admitting that the original decision to accept so much support from the foundation was a mistake. Responding to a question about Templeton's aims, he said: “My sense is that, although they attach no strings [to their money], their interest is to find that great, inevitable position that will bring science and religion together.

That is totally wrong in my view. But what's done is done.” Gould added that the programme was “now on the right path”.

Colin Macilwain



Japan set to tighten ethics rules for genetic sampling

Tokyo

Researchers working on human genetics in Japan will be required for the first time to obtain the informed consent of those providing samples, under draft ethics guidelines released by the Ministry of Health and Welfare last week.

But several leading geneticists have already criticized the guidelines as not doing enough to reassure research subjects that their genetic information will not be misused.

The guidelines are a response to a large project, funded by the ministry, to collect data on single nucleotide polymorphisms (SNPs), due to begin in April. They also follow newspaper articles criticizing various institutions for doing genetic research on human samples without the proper consent of those from whom they were obtained.

The guidelines apply only to research funded by the health ministry. If they are broken, the directors of the institute or hospital in question will be able to stop the project and confiscate its funding.

An ethics panel at the Council for Science and Technology is drawing up similar guidelines. The council advises the Science and Technology Agency — the main funding body for human-genome research in Japan.

Earlier this month, researchers at the National Cardiovascular Center in Osaka, the University of Kyushu, Fukuoka, and Tohoku University, Sendai, admitted that they have used thousands of blood samples for analysis without informing patients.

The revelations have prompted an intense media debate, and observers say that the scientists may have gone public in anticipation of the guidelines. These specify that it is up to institutional ethics boards to decide on the use of samples obtained in the past and without permission.

Scientists at Tohoku University, who halted a research project on the genetics of cerebral apoplexy last autumn to consult the university's ethics board and explain the situation to concerned patients, have declared that they intend to resume research in the next few weeks.

But critics of the guidelines argue that rules on informed consent will not convince patients that their privacy will be protected — particularly as Japan lacks data-protection laws.

“The focus is mainly on how to obtain and how to use human material,” says

▶ Yusuke Nakamura, who directs a large SNPs project at Tokyo University's Institute of Medical Science. "But it is unclear how to ensure the rights of patients, and how to deal with patients whose privacy has been violated, and who may be subject to various forms of genetic discrimination."

Nakamura says that many patients are afraid to provide samples, following speculation that genetics research could lead to social discrimination. "Legal regulation, including penalties for those who do not comply, should be established as soon as possible," he says.

Some observers fear that insufficient regulation could provoke further media criticism, making it harder to obtain human samples. The supply of these is hindered by a lack of links between scientists and clinicians.

"Much recent trouble concerning human samples has been caused by a lack of communication between medical doctors and researchers," says Yoshihide Hayashizaki, who leads the Genome Exploration Research Group at RIKEN's Genomic Sciences Center.

Nobuyoshi Shimizu, who directs the Center for Genomic Medicine at the Keio University School of Medicine, says that it is "difficult to obtain fresh tissue samples". He argues that national facilities are needed to prevent the commercialization of the supply of human tissue samples.

Guidelines for the use of human cells and tissues were published last year by an ethics panel at the Japan Tissue Culture Association, headed by Toshiharu Matsumura, a general manager at the Meiji Cell Technology Center. Matsumura sees them as a "first step" towards a regulatory framework.

Robert Triendi

Gene therapy institute denies that errors led to trial death

Washington

The director of the gene therapy institute where a patient died during treatment last September has admitted errors in the administration of its trials, but denies that they led to the patient's death.

The US Food and Drug Administration (FDA) recently shut down five clinical trials at the University of Pennsylvania's Institute of Human Gene Therapy (IHGT) after an investigation found 18 possible violations in the way the institute had run and monitored an earlier trial (see *Nature* 403, 354; 2000).

In a letter to the FDA, James Wilson, director of the institute, acknowledges errors in the trial — the first in which a death has been attributed directly to the experimental treatment (see *Nature* 401, 517; 1999).

Wilson does not accept all 18 of the FDA's preliminary findings, but he admits perhaps the most serious charge: that the institute failed to contact the FDA when two patients who received an adenoviral vector with a therapeutic gene for a liver-enzyme deficiency experienced adverse effects.

Under the protocol, the IHGT should have notified the FDA immediately. Wilson says that the institute had reported similar effects in two previous patients, and that the FDA had allowed the trial to resume in both cases. He points out that the FDA received written records of those adverse effects before permitting the IHGT to use higher doses on other patients, but wrote that "IHGT should not have proceeded as it did".

But Wilson denies several other serious charges. For example, he says that Jesse Gelsinger, the Arizona man who died in the



Wilson: admits that procedures were lax.

experiment, was eligible for treatment despite having a high level of ammonia before the treatment. The protocol specified only that the enzyme level should be normal at the time of enrolment, not treatment, and it is "common" for such enzyme levels

to fluctuate, Wilson says in his letter.

Also, he says that the researchers did not inform Gelsinger of the deaths of two primates because they were models for a different disease and received a different therapeutic gene.

Finally, the university report contradicts Paul Gelsinger, Jesse's father, who, at a recent Senate hearing, said that investigators did not follow procedures for informed consent properly (see *Nature* 403, 583; 2000). "Each and every patient in the ... trial gave clear and unambiguous consent to participate."

Wilson declined to comment further. But Kenneth Wildes, a university spokesman, says: "We made some mistakes, but those mistakes did not lead to the death of Jesse Gelsinger." Wildes adds: "We're not minimizing the conduct of researchers at Penn. They should be held to the highest standards."

Wilson says in his letter that, if allowed to resume its suspended trials, IHGT would be willing to be more closely monitored, and to follow more stringent reporting requirements.

Paul Smaglik

US physics society puts feisty newsletter in doubt

Washington

What's New, a popular and often scurrilous weekly news summary read by more than 10,000 physicists and followers of science policy in the United States and abroad, could cease publication in July, following the American Physical Society's (APS) threat to stop supporting its author, Bob Park.

The APS wrote last month to the physics department at the University of Maryland at College Park, where Park is a professor, saying that it wants to cut its support from two-thirds to one-half of his salary.

If that happens, says Park, he will have to increase his teaching and other commitments at the university, and will be unable to continue his work for the APS, which

includes the production of *What's New*. Given the lack of effort involved, he describes the APS letter as "an adiabatic firing".

Park has served as a kind of freelance mouthpiece for the APS for many years. He writes frequent columns for *The New York Times* and other newspapers on everything from global warming to the recent spying scandal at nuclear weapons laboratories.

Neither the columns nor the *What's New* electronic newsletter are attributed directly to the APS — the newsletter always ends with Park's signature line: "Opinions are the author's and are not necessarily shared by the APS, but they should be." But they do reflect the views of many physicists, and give the community a rapid response to current events

— lacking in most scientific societies, which can take months to agree a point of view.

As news of the APS plan leaked out during the meeting of the American Association for the Advancement of Science in Washington last week, some physicists were planning to write to the new APS president, Jim Langer, of the University of California at Santa Barbara, to protest at the decision. Some suggest that Park may have offended someone once too often.

Jerome Friedman of the Massachusetts Institute of Technology, last year's APS president, said that he didn't want to discuss the matter "until it was resolved". But he added: "There is no dissatisfaction with Bob — it's a question of resources."

Colin Macilwain

Germany holds up cultivation of GM maize...

Munich

The German government last week unexpectedly stepped in to block the commercial cultivation of genetically modified (GM) *Bt* maize by the biotechnology company Novartis. Citing the need to respect the 'precautionary principle', it is calling for more research into its potential hazards.

The ministers of health, agriculture, environment and research jointly called for a moratorium on the commercial growth of the maize — which has been genetically modified to express insecticidal toxins from the bacterium *Bacillus thuringiensis* — only one day before it was expected to be approved by the ministry of agriculture's Office for Varieties.

Health minister and Green Party member Andrea Fischer instructed the Robert Koch Institute in Berlin, the health ministry's office that licenses field trials, to revoke a licence it had issued to Novartis three years ago. Fischer argued that issues related to the crop's safety have not been resolved.

A report commissioned by Fischer from the Öko Institut, an independent environmental research institute in Freiburg known for its critical stance towards agricultural biotechnology, said it "could not be ruled out" that an antibiotic-resistance marker in the *Bt* maize gene construct would spread antibiotic resistance to the environment. The precautionary principle "must apply", she said. "More research needs to be done."

As a result of the government's move, the Office for Varieties has now put on hold its decision on licensing Novartis's *Bt* maize as a



Genetic marker: Greenpeace campaigns have kept GM maize on the political agenda in Germany.

new variety. Such a licence is required for any new plant variety that is to be cultivated on a large scale.

Environmental groups, including Greenpeace and the German group BUND, have welcomed the decision. But the Organization of German Plant Breeders calls it "legally doubtful and politically questionable".

The German Association of Biotechnology Industries calls it a "depressing decision from the government for the whole of the biotechnology industry". Novartis says it has not decided whether to fight the decision.

The German government's move is in line with the European Commission's *de facto* moratorium on licensing the commercial planting of GM plants.

However, the cultivation and commercialization of the *Bt* maize within the Euro-

pean Union had been authorized in 1997 under rules laid down in a 1991 EU directive.

This authorization has been challenged by some countries. Last year Austria, with the support of Italy and Luxembourg, invoked a clause in the directive allowing authorization to be temporarily suspended at a national level if a member state claimed to have new evidence on safety.

Austria cited health and environmental concerns, particularly fears that the antibiotic-resistance marker in the *Bt* maize gene construct could spread resistance to humans or animals via gut bacterial flora.

The European Commission's scientific advisory committee rejected Austria's evidence on the grounds that it presented no new data. But it has not prosecuted Austria for flouting the directive, given the public's lack of acceptance of GM crops and food.

The Robert Koch Institute is preparing a report for the commission's scientific advisory committee citing similar concerns.

Officials informally admit that the report is unlikely to bring up new facts, other than research results published last December showing that insecticidal concentrations of *Bt* toxin are released into the soil from the roots of *Bt* maize (see *Nature* 402, 480; 1999).

The European Commission is hoping that revisions to its directive, under which authorizations for GM crops would in future be given for only ten years, with continual monitoring for environmental effects, will calm public concerns. The second reading of the revised directive is scheduled for April.

Hans-Günther Gassen, professor of biochemistry at the University of Darmstadt, and spokesman for biotechnology in the science committee of the state of Hessen's ministry of trade, says it is "very difficult to have effective, non-emotional dialogue with the public about agricultural biotechnology, particularly in Germany". His own institute was set on fire by an anti-biotechnology protest group in 1990.

Alison Abbott

... amid calls for international biotech panel

London

Proposals to create an international panel to study the scientific basis of the potential dangers of genetically modified (GM) crops — comparable to that which already exists for climate change — could emerge from an intergovernmental meeting to be held in Edinburgh next week.

The invitation-only meeting will involve about 400 scientists, politicians, policy-makers and non-governmental organizations. It has been organized by the Paris-based Organization for Economic Cooperation and Development at the request of heads of the so-called G8 group of industrialized nations, in particular the French president Jacques

Chirac, under the title 'GM Food Safety: Facts, Uncertainties and Assessment'.

Sir John Krebs, chairman of the conference and recently appointed head of Britain's Food Standards Agency, says that one idea put forward by "a number of people" is an international forum for debating scientific issues, like the International Panel on Climate Change.

"If one went forward with a panel that looked at the science, one would want it to be a scientifically focused forum, but also to recognize that one would then have to fold in value systems, beliefs, economic implications, and so on," says Krebs, who is also professor of zoology at the

University of Oxford.

Krebs says that the main purpose of the meeting is to identify, through dialogue between individuals and groups with different, and often opposing, ideas, "areas of greater convergence and areas of lesser convergence".

"What we have seen in the past few years is a slanging match. I would hope that this might trigger a long-term process by which other matters related to trade and the environment can be debated in an open way," says Krebs.

A report on the Edinburgh meeting is due to be delivered when the G8 industrialized nations next meet.

David Dickson

Energy department labs face threat to IT research funds

Washington The US House of Representatives last week passed a research bill that would take \$300 million out of the Department of Energy's budget for information-technology (IT) research over five years — half of the total they currently plan to spend.

The money would be transferred from the department's labs to the National Science Foundation through an unopposed amendment attached to the bill by Michael Capuano (Democrat, Massachusetts). Congressional staff say that Capuano is angry about the department's lack of responsiveness to his concerns over the unrelated matter of heating-fuel costs.

The transfer is unlikely to become law, as the Senate has yet to consider a companion measure. But the move by the House will damage the labs' chances of getting a share of money for research in advanced computing.

Pope ignores the masses over 'heretic' astronomer

Rome A demonstration in the Campo di Fiori in Rome last week marked the four-hundredth anniversary of Giordano Bruno's

burning at the stake — probably in the same square — for claiming that there is an infinity of worlds. But Pope John Paul II is unlikely to include Bruno in his 12 March 'day of pardon' to mark the millennium.

Speaking on behalf of the Pope, Cardinal Angelo Sodana, secretary of state to the Vatican, told a meeting in Naples last week that the Vatican would reconsider the Bruno case "with a spirit open to the full historical truth". But he added that Bruno's philosophy "remains incompatible with orthodox Christian thinking".

Siberian accelerator up to speed

Moscow A new electron-positron accelerator has begun operation at the Russian Academy of Sciences' Institute of Nuclear Physics in Novosibirsk, Siberia. "There are still a few sectors in nuclear physics where Russian scientists are at a world level," says Aleksandr Skrinsky, director of the institute. "One of these is the investigation of the strong interactions by using colliding beams of electrons and positrons."

Although the Russian beamline has a circumference of only 300 m, compared with 27 km at the European Laboratory for Particle Physics (CERN) in Geneva, Skrinsky argues that Russian research is not

lagging behind. The first accelerator using colliding electron-positron beams was built in Novosibirsk in 1965.

French environmental labs may get 'big science' funds

Paris France is considering expanding its network of environmental and ecological laboratories to study the long-term impact of environmental change on populations, communities and ecosystems. The stations would also contribute to the international effort to collect information about the biosphere.

Alain Pavé, former director of the environment, life and societies programme at the Centre National de la Recherche Scientifique, has proposed that the government fund this effort through its 'big science' budget. The government is considering whether to expand this budget — which funds facilities such as particle accelerators, X-ray sources and observatories — to include projects in biology and the environment.

AAAS honours Russian for nuclear campaign...

Washington Alexandr Nikitin, a former engineer arrested for reporting on radioactive contamination in Russia, has been given the

scientific freedom and responsibility award by the American Association for the Advancement of Science (AAAS) at its annual meeting in Washington. But Nikitin, who was acquitted of divulging state secrets last December, could not collect the award as the Russian security police have not released his passport.

Nikitin has been under investigation since 1995, and was arrested and held for ten months in 1996. But the St Petersburg prosecutor has appealed against the verdict, and has called for the ruling to be dismissed.

... and names botanist Raven as next president

Washington Botanist Peter Raven was made president-elect of the American Association for the Advancement of Science at its annual meeting in Washington this week. Raven is director of the Missouri Botanical Garden in St Louis and a professor at Washington University in St Louis.

At the meeting, Mary Good, a chemistry professor at the University of Arkansas, Little Rock, assumed the presidency of the association. Good has served in scientific posts for four consecutive US presidents and is the managing member for the company Venture Capital Investors. She replaces Stephen Jay Gould of Harvard University.

European bodies call for faster networks

Paris High-speed research networks are needed to underpin radical new approaches to research, according to a joint statement issued by the European Science Foundation and Academia Europaea last week. The statement calls on European institutions and governments to promote a new generation of networks running thousands of times faster than those used today.

Europe lags badly behind the United States in this area, the statement says. "If the state of affairs is allowed to continue," it adds, "Europe will be left out of major advances in research and education." But the political difficulties among European nations of agreeing on big shared infrastructure hindered the creation of a counterpart to US initiatives, according to the report (<http://www.esf.org/ftp/pdf/SciencePolicy/ESPB7.pdf>).

UN brings in German rector to run University of Kosovo

Munich Michael Daxner, former rector of the University of Oldenburg in Germany, begins a two-year stint this week as interim administrator of the university system in Kosovo, as part of a bid by the United Nations' Security

Council to stabilize and rebuild the region after the war.

The University of Kosovo, torn apart by the ethnic violence, is estimated to have 1,600 faculty members — none of whom currently have contracts — and around 20,000 students. Daxner says his first task will be to create a new legal and financial framework for the university, designed along western European lines. "Most important is the creation of a single entity, with an ethnic mix of Albanian and Serbian Kosovars," he says.

Venter and Wilson share King Faisal prize

London Craig Venter, head of Celera Genomics, and Edward Wilson of Harvard University have won the 2000 King Faisal prize for science, awarded this year in biology. The prize includes a medal and a shared cash award of US\$200,000.

Venter was cited for his novel techniques for the rapid identification of genes, the fast and economical sequencing of genomes, and his "significant contribution to the elucidation of the human genome".

Wilson was described as one of the most outstanding biologists of the century and a pioneer in sociobiology, the study of species within ecosystems, and the conservation of biodiversity.

Open annotation offers a democratic solution to genome sequencing

Sir—Jean-Michel Claverie¹ writes in Correspondence about the problems of annotating the whole human genome sequence, given that a draft form will be available in a few months. While we agree with many of his points, we disagree with what he says about the lack of bioinformatics capacity to provide a useful basic analysis. The Sanger laboratories, with the European Molecular Biology Laboratory's European Bioinformatics Institute, have been developing an automatic analysis system for some months; the results of the first full release of Ensembl can be seen at <http://www.ensembl.org/>. The system now tracks the daily output of human genomic sequence in real time. It is based on confirming *ab initio* predictions by homology and providing functional annotation via Pfam². So far 17,045 gene fragments are annotated from the 1,405,539,258 bases processed.

We agree with Claverie about the limitations of any automatic analysis system, having ourselves worked on the semi-manual analysis of the human chromosome 22 sequence. However, a large subset of genes can already be predicted accurately, which will be very useful as a way into this huge volume of data. A key aspect of the system is its ability to keep track of genes despite revisions to the sequence. This will be important as the genome is completely sequenced over the next couple of years. Ensembl accession numbers assigned to genes are permanent identifiers that will refer to the same genes throughout this process.

How can we go beyond this baseline automatic annotation? Claverie points out the chaos that would result from duplicated annotation efforts, each with different standards and different ways of presenting the data. He is also correct in arguing that no single collaborative group will be capable of annotating the entire genome consistently and to high quality. One way to deal with this is to have a monolithic single entity that invests 300 person-years into annotating the genome. A better one is 'open annotation', where the annotation required is distributed across a highly motivated community of biologists.

We believe that many of the problems with open annotation are technical ones, which can be and are being addressed. The web allows different data sources to be readily crosslinked, but different websites have different formats and interfaces. An alternative, particularly appropriate for sequence data, is for a browser to merge

annotation from multiple data sources on top of a baseline coordinate system to provide the user with a single annotation view. Lincoln Stein and colleagues are developing such a system (DAS) based on XML (see <http://stein.cshl.org/das/>). All that is then required for any centre to contribute annotation of all or part of the genome is to synchronize its coordinate system with its baseline server. Maintaining the coordinate system across a changing genome does require substantial resources, but keeping in synchronization with this need not. Ensembl is an open-source project and will provide both a common object framework for annotation as well as the synchronization tools needed for anyone to set up to serve annotation for all to see and use.

The power of open-source software is well recognized³, although it could be feared that open annotation will swamp biologists with alternative contradictory views of the sequence. We are more optimistic. Browsers will allow biologists to select only the data sources they wish to view. Just as some websites become popular, word of useful annotation will spread quickly, since selecting it will be as easy as bookmarking a new website. Software development has been democratized by open-source projects such as Linux, which have allowed everyone the opportunity to contribute. Open annotation provides the same opportunity for genomes, and so should speed our collective decoding of genetics without centralized annotation centres or commercial monopolies.

Tim Hubbard*, Ewan Birney†

*Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK

†EMBL European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK

1. Claverie, J.-M. *Nature* **403**, 12 (2000).

2. Bateman, A. *et al. Nucleic Acids Res.* **28**, 263–266 (2000).

3. *Nature* **403**, 229 (2000).

Affirmative action won't solve sex discrimination

Sir—Natasha Loder's article and the accompanying cartoon on gender discrimination (*Nature* **402**, 337; 1999) shed little light on this vexed issue. Some people have no doubt that discrimination, sexual and otherwise, does exist in academic institutions, as it does in most other human endeavours. Others consider this merely a reasonable working hypothesis requiring clear evidence, the type of evidence that the European Technology Assessment Network report attempts to provide.

The excerpt from the report presented in the article, concerning the low proportion of women in national scientific academies, is unconvincing to anyone, male or female. It is clear that there are fewer women in the upper echelons of academic research, but there are many possible reasons. The most parsimonious of these is that the long climb up the academic ladder takes a few decades, and the present demographics in national scientific academies reflect newly trained scientists emerging 20 or 30 years ago.

If we accept that gender-based discrimination is wrong, we should at least try to examine the problem more rigorously before suggesting sexually discriminatory policies aimed at ensuring a gender balance on public bodies. If, indeed, time-lags are responsible for the gender disparities, they will disappear in due course, independently of changes in hiring and funding practices. Discrimination is the problem, not the solution.

G. A. Lozano

Department of Biological Sciences, Simon Fraser University, Burnaby, BC V5A 1S6, Canada

Why Roche deserves the disputed *Taq* patent

Sir—You report¹ the legal decision in the long-running Roche–Promega dispute, that the US patent on full-length *Taq* DNA polymerase is invalid. The patent (4,889,818) had been awarded to Cetus Corporation and subsequently bought by Roche. This decision does not affect the validity or enforceability of Roche's foundational patents on polymerase chain reaction (PCR) and other related patents, including those for using any thermostable enzyme, including native *Taq*, for PCR.

I believe that the court's decision was wrong and unfair to Cetus scientists David Gelfand and Susanne Stoffel, in that it did not distinguish their invention from the work of the Gorodetskii² and Trela³ groups. Cetus scientists including Gelfand and Stoffel were the first⁴ to isolate and clone the full-length (molecular mass 94,000; 94K) *Taq* DNA polymerase. The earlier groups repeatedly published their isolation of *Taq* fragments (60K–70K), undoubtedly the result of proteolytic degradation, under the mistaken impression that it was the complete enzyme.

Instead of concentrating on the validity of the Cetus invention, Promega's case was based on misrepresenting the raw experimental data of the scientists and their good-faith interpretations of it. By this stratagem, Promega is trying to rewrite history by asserting that Cetus had data indicating that the earlier groups isolated a 94K

enzyme even when those laboratories' own publications show clearly and repeatedly that they did not. Readers can obtain these articles and judge for themselves who deserves credit for isolating and cloning the full-length *Taq* DNA polymerase.

T. J. White

Roche Molecular Systems, 1145 Atlantic Avenue, Alameda, California 94501, USA

1. Dalton, R. *Nature* **402**, 709 (1999).
2. Kaledin, A. S. *et al. Biokhimiya* **45**, 644–651 (1980).
3. Chien, A. *et al. J. Bacteriol.* **127**, 1550–1557 (1976).
4. Lawyer, F. C. *et al. J. Biol. Chem.* **264**, 6427–6437 (1991).

ECT has proved effective in treating depression ...

Sir—Peter Sterling¹ criticizes the reviewer of my book² for stating that electroconvulsive therapy (ECT) is both safe and effective in treating the severely mentally ill³. Using the example of a friend who, a year after having ECT, complained of “huge gaps in recall of major life events”, he argues that history will view ECT as another “great and desperate cure” akin to psychosurgery.

Sterling errs on many counts, the principal one being his reliance on images of ECT from half a century ago. Despite Sterling's assertion, electric currents are not inherently dangerous just because touching an open electric socket is painful. For example, they are widely used to slow a racing heart. In ECT, because of the resistance of skin, bone and galea, very little of the energy gets to brain tissue; indeed, modern technique calls for the determination of the minimum current needed to elicit a seizure.

Scientists have sought evidence of brain damage for half a century, using elaborate brain imaging methods and psychological tests in humans, and neuropathological studies in animals given high levels of currents or large numbers of seizures. No such evidence has emerged⁴.

Sterling is wrong when he argues that the cure for depression “requires this procedure to be repeated 10–20 times over a week or so”. Modern ECT resolves depression in four to eight sessions, spaced at two to three times a week, in more than 80 per cent of patients using optimal treatment methods. No study has found any other antidepressant treatment to be more effective than ECT⁵.

ECT has no semblance to stroke, and Sterling's comparison is unfounded. Again, contrary to Sterling's argument, studies of the functions of the non-dominant hemisphere fail to demonstrate persistent effects of unilateral ECT on both non-verbal and verbal testing^{5,6}.

ECT was developed in the late 1930s,

and by 1950 tests of memory function were part of our treatment schedule^{2,5}. Such tests are routine today, and the literature on ECT and memory is extensive⁶. There are defects in memory associated with severe mental illnesses, and these defects are exaggerated during treatment with medications or ECT. But when patients recover from their illness, and from the immediate effects of ECT or medicines, they are as able to recall life events, to learn anew and to work effectively as age- and education-matched controls.

The benefits of ECT are temporary, but so are the benefits of psychiatric medications. These treatments must be continued for years to prevent relapse.

Max Fink

PO Box 457, St James, New York 11780, USA

1. Sterling, P. *Nature* **403**, 242 (2000).
2. Fink, M. *Electroshock: Restoring the Mind* (Oxford Univ. Press, New York, 1999).
3. Freeman, H. *Nature* **401**, 327 (1999).
4. Devanand, D. P., Dwork, A. J., Hutchinson, E. R., Bolwig, T. G. & Sackeim, H. A. *Am. J. Psychiatry* **151**, 957–970 (1994).
5. Abrams, R. *Electroconvulsive Therapy* 3rd edn (Oxford Univ. Press, New York, 1997).
6. Sackeim, H. A. in *Cognitive Disorders: Pathophysiology and Treatment* (eds Moos, W. H., Gamzu, E. R. & Thal, L. J.) 183–228 (Dekker, New York, 1992).

... and there's no proof of lasting brain damage

Sir—Peter Sterling¹ asserts that electroconvulsive therapy (ECT) damages the cerebral hemispheres. A comprehensive review of the relevant literature provides no objective evidence that ECT is capable of causing brain damage in human beings^{2,3}.

In support of his belief that psychiatrists would find evidence for ECT-induced brain damage if only they would test for it, Sterling cites a single study, conducted more than 50 years ago. In sharp contrast are the many studies in the modern era examining short- and long-term memory before and after bilateral or unilateral ECT, examining both left- and right-hemisphere functions and including non-ECT control groups for comparison. All but one of these studies reveals no evidence for persistent or long-term deficits².

The single positive study⁴ found detectable deficits in autobiographical memory six months after bilateral (but not unilateral) ECT. Unfortunately, this was published only in academy proceedings, which are not subjected to peer review.

Sterling's claim that ECT “releases massive quantities of glutamate”, thus causing neuronal death, is not supported by any citation. However, myelin basic protein, which is found in the cerebrospinal fluid following stroke, does not

appear there following ECT⁵, nor does the CPK brain isoenzyme⁶.

Richard Abrams

Department of Psychiatry, Chicago Medical School, 3333 Green Bay Road, North Chicago, Illinois 60064, USA

Director of Somatics, Inc., manufacturer of the Thymatron ECT device

1. Sterling, P. *Nature* **403**, 242 (2000).
2. Abrams, R. *Electroconvulsive Therapy* 3rd edn (Oxford Univ. Press, New York, 1997).
3. Devanand, D. P., Dwork, A. J., Hutchinson, E. R., Bolwig, T. G. & Sackeim, H. A. *Am. J. Psychiatry* **151**, 957–970 (1994).
4. Weiner, R. D., Rogers, H. J. & Davidson, J. R. T. *et al. Ann. NY Acad. Sci.* **462**, 315–325 (1986).
5. Hoyle, N. R. *et al. Br. Med. J.* **288**, 1110–1111 (1984).
6. Dec, G. W. *et al. J. Am. Med. Assoc.* **253**, 2525–2529 (1985).

Shrinking shrews

Sir—Wikelski and Thom, in their interesting description of marine iguanas *Amblyrhynchus cristatus* shrinking in response to food shortage¹, thought theirs was the first report of shrinkage in adult vertebrates. But in 1949, the Polish mammalogist August Dehnel drew attention to the shrinkage overwinter of braincase size in common shrews *Sorex araneus*². This phenomenon, which now bears Dehnel's name, affects other organs (especially the liver and kidneys), and occurs in all northern soricines examined so far. It is, like the changes in the marine iguanas, a reversible response of individuals, not just a phenomenon of populations³.

D. W. Yalden

School of Biological Sciences, University of Manchester, 3.239 Stopford Building, Oxford Road, Manchester M13 9PL, UK

1. Wikelski, M. & Thom, C. *Nature* **403**, 37 (2000).
2. Dehnel, A. *Ann. Univ. M. Curie-Sklod.* **4**, 18–102 (1949).
3. Churchfield, S. *The Natural History of Shrews* (Christopher Helm, London, 1990).

Cloning shock: was Dolly a three-legged accident?

Sir—We all know Dolly was cloned from one of her dam's cells, but why does *Nature* publish a photo of her¹ with three of her own legs and one of her mother's apparently grafted on to her umbilicus? On the cover of the *Nature* issue that announced Dolly, this fourth leg had a black hoof, quite properly since—in the photo² from which the cover image was taken—it belonged to the Scottish Blackface dam standing behind her. It's time the world was told what happened to Dolly's other leg.

Wendy Gibson

School of Biological Sciences, University of Bristol, Woodland Road, Bristol BS8 1UG, UK

1. *Nature* **403**, 479 (2000).
2. *Nature* **385**, 812 (1997).

A picaresque genius

Contributions of a flamboyant scientist, businessman and self-aggrandizer.

Félix d'Herelle and the Origins of Molecular Biology

by William C. Summers

Yale University Press: 1999. 230 pp. \$30, £20

Gunther S. Stent

Bacterial viruses were discovered in 1915 by the British bacteriologist Frederick Twort, whose publication of his findings in *The Lancet* went completely unnoticed. Twort was rescued from obscurity several years later, in the aftermath of an entirely analogous discovery made in 1917 by Félix d'Herelle, a French-Canadian bacteriologist, who never acknowledged Twort's priority. D'Herelle's announcement of these agents — he called them "bacteriophages" — produced a sensation in the world of medical microbiology. He claimed that they, rather than antibodies in the blood serum, were the chief mediators of natural immunity against bacterial diseases, and promised that they would provide the means for a universal prophylaxis and therapy.

He was mistaken. Yet although d'Herelle's flamboyant personality kept him under constant attack from his colleagues, even they thought he was right. In fact, they complained that these claims were so obviously true that he deserved little credit for making them. Their fiercest disdain was directed at his proposition that the bacteriophages, whose basic life cycle he had managed to establish within three years of their discovery, were self-reproducing viruses living parasitically off bacteria. In that, however, d'Herelle was not mistaken.

He showed that the parental virus particle attaches itself to the surface of a bacterial host cell, penetrates it, and reproduces itself on the inside to generate a clone of many progeny particles within half an hour. When the infected host bacterium dissolves, the progeny virus particles are released into the medium, ready to infect and kill further bacteria for another reproductive cycle.

Although d'Herelle's bacteriophages failed to produce a medical panacea, they did become one of the principal experimental materials of the fledgling discipline of molecular biology, which, by the end of the twentieth century, had totally transformed the life sciences.

This happened when the German physicist Max Delbrück ran across bacterial viruses by chance at Caltech in 1937 as a young postdoctoral fellow. Appreciating the possibilities d'Herelle's bacteriophages offered for carrying out quantitatively precise experiments with a rapidly self-reproducing organism, Delbrück adopted



Acknowledged but reviled: d'Herelle's talents included making enemies of powerful scientists.

them as his experimental material.

The central problem of life that Delbrück wanted to address was: how does life manage to beget like? He boiled it down to the question: how does the parental bacterial-virus particle manage to replicate itself inside its bacterial host cell to give rise to its multitudinous brood of genetically identical progeny? In 1940, he founded the American Phage Group, which not only came to dominate bacterial-virus research, but, more importantly, exerted a decisive influence on the development of the then still-unnamed discipline of molecular biology.

In *Félix d'Herelle and the Origins of Molecular Biology*, the first full-length biography, d'Herelle emerges as one of the most picaresque geniuses of twentieth-century

biology. Its author, the Yale University biophysicist William C. Summers, drew on personal interviews with people who knew d'Herelle as well as on published primary and secondary sources. He also made a surprising discovery of his own: an unpublished, previously unknown autobiographical memoir by his subject.

As a member of the American Phage Group, I considered myself an expert on the history of the bacteriophage, but this painstakingly researched account taught me much about d'Herelle. An internationally famous scientist and a credible Nobel prize candidate, d'Herelle had no formal education beyond graduation from a Parisian *lycée*. He trained himself as a microbiologist in his private home laboratory before embarking on a globe-trotting career. The microbiological projects he pursued ranged from routine bacteriological examinations in public-health facilities to the development of yeast strains to make cheap whisky from rotting tropical fruit.

His successful microbiological control of locust plagues set the pattern for later attempts to suppress insect pests by using biological agents rather than chemical insecticides. It also set the pattern for d'Herelle's future standing in the scientific community. He became widely known for his imaginative approaches to important problems in theoretical, as well as applied, microbiology. At the same time, he was widely reviled for his self-advertisement, his exaggerated claims of success and his sharp financial practices. He also had a talent for making enemies among powerful senior scientists.

At the Pasteur Institute, where he began working as a volunteer in 1915, his enemies managed to prevent his appointment to a regular staff position. Yet it was when the institute's director sent him to investigate an



Cultivating success: an advertisement for phage from d'Herelle's commercial laboratory.

book reviews

epidemic of dysentery that was raging in a cavalry squadron stationed in a Paris suburb that he discovered bacterial viruses.

D'Herelle had filtered an emulsion from the sick men's faeces, added a drop of the bacterium-free filtrate to a culture of dysentery bacteria and spread the mixture on a solid surface of nutritive jelly. After incubating this overnight, he found a dense lawn of bacteria had grown that — unlike the continuous lawns formed by normal bacteria — contained a few circular clear spots. He called these "plaques" and correctly inferred that each plaque represents an area where the bacteria had been destroyed by the clone of progeny descended from a single parental virus particle. On the basis of this insight, d'Herelle developed his plaque method of assaying the number of infective virus particles in a given volume of solution, opening the door to quantitative virology.

His first regular academic job came in

1922, when the Institute of Tropical Medicine at Leiden University in Holland appointed him as conservator and awarded him an honorary MD degree. On the basis of this he styled himself "Doctor d'Herelle" — so when the Egyptian Suez Canal Authority hired him as a quarantine officer in Alexandria during 1924–27, it believed him to be a bona fide medical doctor. And when he made the academic big time in 1928, as a professor of bacteriology at Yale, the university was unaware that he owned and actively managed a commercial laboratory in Paris producing bacteriophage preparations for sale to the pharmaceutical trade. He spent hardly any time in New Haven and made an enemy of the dean of the medical school. By 1933, his relations with Yale had deteriorated so badly that he resigned.

At his next job, at the Bacteriophage Research Institute in Tiflis, Soviet Georgia, d'Herelle made an attempt to be politic. The

1935 Russian edition of his book *Bacteriophage and the Phenomenon of Recovery*, published in Tiflis in 1935, was dedicated "to him who possesses the ruthless and relentless logic of history, and who stands at the threshold of society; to him who has reached the top: COMRADE STALIN". Two years later, Comrade Stalin had the director of the Tiflis institute executed for alleged counter-revolutionary activities, and d'Herelle returned to Paris.

The biography is edifying, but its title is misleading. D'Herelle had nothing to do with the conceptual or ideological origins of molecular biology; he merely isolated one of its paradigmatic experimental organisms. However, it is true to the spirit of d'Herelle, who dearly enjoyed accepting undeserved credit.

Gunther S. Stent is in the Department of Molecular and Cell Biology, University of California, Berkeley, California 94708-3200, USA.



Enter the cosmic cathedral

The Rose Center for Earth and Space Science, New York
P. M. Solomon

The new Rose Center for Earth and Space Science at the American Museum of Natural History in New York opened to the public on 19 February. The \$210 million structure is a glorious, 120-foot-high, glass-enclosed hall containing a sphere 87 feet in diameter as its centrepiece. The sphere, which appears to be suspended in space, contains the completely new Hayden Planetarium. This replaces the famous original, which served millions of visitors for 65 years.

Viewed from the outside, scale models of Jupiter, Saturn and the other planets appear in front of the sphere, which represents the Sun. The centre, designed by Polshek Partnership Architects as a "cosmic cathedral", is a stunning addition to the public architecture of New York City, rivalling the Guggenheim Museum on the other side of Central Park. Under the leadership of the museum's president, Ellen Futter, construction was completed in under three years, to be ready to open at the start of the new millennium.

The director of the Hayden Planetarium is astronomer Neil de Grasse Tyson, who was appointed in 1996. In the past year he has been joined by a new eight-member Department of Astrophysics headed by Michael Shara. The museum has clearly made a major commitment to education

and research in astronomy. The Rose Center is a combined public and private effort, with major funding from many donors, including Frederick P. and Sandra Rose, New York City and New York State.

In addition to the Hayden Planetarium, the Rose Center has many major exhibits in the lobby-level Hall of the Universe, together with an elaborate display of *Scales of the Universe* and a *Cosmic Pathway*. The displays are designed to convey to visitors a sense of time, space and distance in the Universe, along with a coherent picture of the concepts that govern the formation and evolution of planets, stars and galaxies. Basic processes such as nucleosynthesis and the expansion of the Universe are also covered. The exhibits are integrated with the architecture and use the latest technology, including three-dimensional images and interactive displays.

Along a second-floor walkway traversing the glass wall is the *Scales of the Universe* exhibit. This explores the immense range of size in the cosmos using powers of ten incrementally with models and text to show relative sizes, from clusters of galaxies down to stars, planets, cells and atoms. From the walkway, the suspended planet models and the sphere of the Hall of the Universe are clearly visible.

The *Cosmic Pathway* is an analogue walk through the Universe along a 360-foot-long spiral ramp. It starts with the Big Bang at time zero and ends at the present, 13 billion years later. Visitors can measure their stride in millions of years. The pathway contains 220 astronomical images of objects observed at distances that correspond to the age of the Universe when their light was emitted, starting with the furthest-identified galaxy at a redshift of six. Thirteen markers along the way indicate the passage



India's scientific sculpture: eighteenth-century sundials at the Jantar Mantar observatory in Jaipur.

of each billion years. Events such as the formation of the Milky Way and of the Solar System are indicated along the pathway, and at the very end the thickness of a human hair dramatically illustrates the brief relative duration of human history since the era of cave paintings.

Fortunately, the exhibit can be kept up to date, since astronomers' estimates for the precise age of the Universe depend on the Hubble constant and the determination of the deceleration or acceleration of the expansion of the Universe, numbers that change frequently as the latest measurements are reported.

The most spectacular exhibit is the *Passport to the Universe* show in the 420-seat Space Theater located in the upper hemisphere of the Hayden Planetarium. The theatre is equipped with the world's largest and most powerful virtual-reality simulator, which takes the visitor on a flight through the Universe based on real astronomical data and numerical models, accompanied by music and an entertaining and informative narration. A view of the sky as seen from Earth is generated by a Zeiss Mark IX star projector built to the museum's specifications, utilizing fibre-optics imaging to project as many as 9,100 stars onto the dome along with the Sun and planets.

At the core of the virtual Universe is a digital galaxy which includes astronomical data from catalogues of stars, nebulae, interstellar gas clouds and galaxies, including the entire list of the 100,000 observed stars nearest to Earth compiled by the Hipparcos Satellite. The billions of uncatalogued stars in the rest of the Milky Way are represented by a statistical simulation based on galactic models to recreate a three-dimensional volume of objects within 100,000 light years of

the Sun. The entire database is stored on a silicon-graphics supercomputer visual workstation, which can update the location of all objects 30 times per second as the viewer travels through space. The output is projected onto the dome by seven projectors with a total of 7 million pixels.

The technology and software developed by the museum staff in collaboration with National Science Foundation Supercomputer Centers and simulation-software engineers works extremely well, and the virtual voyage through the Universe is breathtaking. Highlights include a flyby of Jupiter and its moons; travelling directly beneath Saturn and its rings; through our local neighbourhood of the Milky Way into the Orion Nebula; out of the Milky Way and through intergalactic space at millions of light years per second; past nearby galaxies, to part of the local supercluster of galaxies; and beyond to the distant Universe represented by hydrodynamical models. The voyage through the colourful Orion Nebula, based on imaging of this active star-formation region by the Hubble Space Telescope, is particularly impressive in conveying detail and complex three-dimensional structure. The show will delight and educate visitors of all ages and expertise from all over the world, including the half million schoolchildren every year who visit the Hayden Planetarium.

An excellent companion book which covers a very wide range of topics, *One Universe: At Home in the Cosmos* (Joseph Henry Press, \$40), brings modern astronomy to the lay reader with a combination of exceptional illustrations, photographs and extensive earthly analogies. ■

P. M. Solomon is in the Department of Physics and Astronomy, SUNY at Stony Brook, Stony Brook, New York 11794, USA.

India's fruitful tryst with technology

Innovative India

edited by L. K. Sharma and Sima Sharma
Medialand: 1999. 352 pp. £48

V. S. Arunachalam

More than 50 years ago, when India gained independence, Prime Minister Jawaharlal Nehru spoke of the country's tryst with destiny. He could as well have spoken of India's tryst with the world of science and technology. For his successors have continued his support for these pursuits, though not often in cash. No other developing country, not even China, has articulated such a passion for science and technology as has India.

The Indian record is truly impressive. From ancient times, the country has periodically yielded unique insights, including the concept of zero, which the Indians are never tired of recounting, and of infinity. In the nineteenth century, India took to occidental education and to modern science. At first, scientific research was carried out at a few metropolitan universities, but it soon spread to the vast conglomeration of laboratories created by a newly independent India.

Innovative India, a glossy coffee-table publication, recounts the recent history of Indian science and technology with articles by practically all the major names in the field, plus a few by foreign friends of the Indian scientific community. It is edited by husband-and-wife team L. K. and Sima Sharma: he is a veteran journalist specializing in Indian foreign policy and science; she is the editor of the prestigious volumes published by the India International Centre. But by allowing the scientists and, in many instances, the bureaucrats to tell their stories, the editors lost a unique opportunity to judge and make their own preferences known. India has many tales of success to choose from, and also of failure: this volume projects all stories, successes and failures alike. Perhaps an integration of this kind is necessary to tell the world how far Indian science and technology has travelled and the meandering routes it has taken. But the reader must be the judge.

There is no better success story to start with than that of the 'Green Revolution', the plant-breeding programme created by American agricultural researcher Norman Borlaug. Ushered in by scientists and educators at Indian agricultural research laboratories and universities, with unstinting technical cooperation from Borlaug and his colleagues in the United States, this allowed India to replace the begging bowl with a bread basket. And what a success: it changed the body politic and freed the country from

dependence on other nations' charity. The Indian space programme, mercifully free from military-missile projects, has brought home the power of satellites to the community in remote sensing, weather forecasting and communications.

The civilian atomic-energy programme is another story. The 13 pages covering this topic do not mention that it has failed to meet even the modest target, set years ago, of 10,000 MW, nor that breeder power reactors (touted as the only answer to India's endemic power shortage) are not realizable for at least five more decades. There are also stories of government science departments that seem to have no other reason for existing — other scientific organizations are doing their jobs better and more cost-effectively — except as retreats for bureaucrats and has-been scientists.

A few universities and, particularly, the Indian Institutes of Technology (IIT) have done exceptionally well. Professor M. M. Sharma, in different sections of this volume, describes the achievements of a single university department in partnering innovations in chemical industries in and around Mumbai. But the IIT's truly outstanding performance in building and sustaining high-quality engineering education and research are achievements that any country could be proud of. Many Silicon Valley entrepreneurs (and millionaires) are graduates from these institutes.

The country's achievements in biotechnology, fuelled by the Biotechnology Board, suggest that this field may soon turn out to be an Indian success story. If, in spite of all these, India remains poor with a high illiteracy rate, we must look for reasons elsewhere: failure to control population growth, or to eradicate illiteracy and eliminate the country's endemic corruption?

In the nineteenth century, Lord Thomas Macaulay argued for introduction of English as the language of teaching in India because he anticipated the need for a large number of clerks and petty bureaucrats to serve the Empire. Indians have every reason to thank him for it — not for the clerks the system produced, but for the scientists. This book recounts the road travelled with feeling. ■

V. S. Arunachalam is in the Departments of Engineering & Public Policy and of Materials Science & Engineering, and at the Robotics Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA.

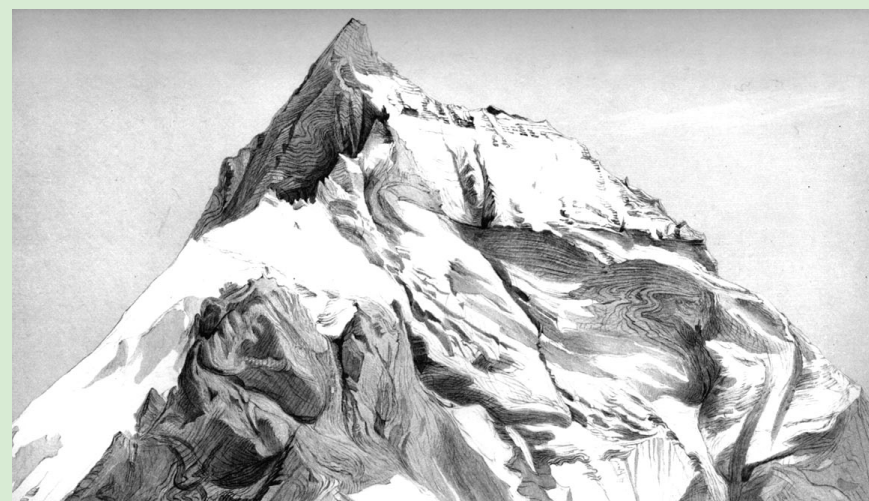
More on science in Asia

Nature and the Orient:

The Environmental History of South and Southeast Asia

edited by Richard H. Grove, Vinita Damodaran & Satpal Sangwan
Oxford University Press, £35

Science in culture



John Ruskin (1819–1900), man of art and science

Douglas Palmer

Our understanding and appreciation of landscape, and our access to some of the most beautiful countryside in Britain owe much to the powers of imagination and descriptive ability of that high priest of Victorian culture, John Ruskin, who died 100 years ago (on 20 January 1900). Although he had a predominantly arts background from his early education, he made serious studies of many areas of contemporary Victorian science, especially botany and geology, and helped to link science and the arts.

Ruskin was the first to anatomize and explain the surface form of landscape for the general reader, especially in volume IV of *Modern Painters* (1856). He simply applied the traditional methods of his training as a visual artist. To understand the surface form of landscape — the skin of the Earth — he realized that we must understand its underlying anatomy, in other words, its geological structure and material.

Rocky landscapes had been faithfully observed and accurately sketched by the likes of John Turner, Ruskin's hero among painters, but the underlying geological structure was not known to Turner. For many artists, landscape painting could never be the same again after Ruskin had revealed its hidden truths.

Ruskin looked to the work of professional geologists such as John Phillips, a colleague at Oxford, and especially the work of Horace Benedict de Saussure on the Alps. Ruskin took his geological studies seriously and was elected a Fellow of the Geological Society of London in 1840. Although he never contributed to the society's publications, he did write a number of geological essays for the Cambridge-based *Geological Magazine*.

Ruskin's artistic and descriptive skills allowed him to accurately depict the various rock types of the Alps, their igneous granites, metamorphic marbles, gneisses and schists, with their complex cleavages and folds. From this material base, he

Ruskin's Matterhorn: contorted by folds and faults as if it were made of some plastic clay.

turned his attention to the large-scale structures, the immense, spectacular folds and faults that contort the great rock masses as if they were made of some plastic clay. Although his geological interpretations are somewhat out of date, his wonderfully accurate drawings and paintings clearly show the relationship of the surface topography to the underlying structure and rock type. In volume IV of *Modern Painters*, Ruskin elaborately details his understanding of the geology of the Alps with the forms of the aiguilles (sharp peaks), ridges, precipices and banks.

Ruskin construed all this structure as part of a purposeful design within the teleology of a fairly fundamentalist Christian belief. For many Christian thinkers, there had been a major question over what was the God-ordained use of infertile and wild mountain country. To Ruskin it was "absolutely necessary that such eminences should be created, in order to fit the earth ... for human habitation". Without mountains, air could not be purified, the flow of rivers sustained, and humans could not be startled out of their lethargy with the "deep and pure agitation of astonishment". For Ruskin, the grand and noble architecture of the mountains are the result of a higher mission with a deeply Christian moral agenda. They appeal to the higher faculties of the human spirit.

For scientists today, Ruskin's agenda may not seem particularly interesting. But his ideas and spiritual evangelism about landscape became influential in his day through his popular writing. His response to landscape led to his belief that access to mountain scenery was a spiritual necessity for Britain's urban poor. The practical impact of Ruskin's Christian socialism on Octavia Hill and Hardwicke Drummond Rawnsley, two of the founding members of Britain's National Trust, was particularly important.

Douglas Palmer is at 31 Mawson Road, Cambridge CB1 2DZ, UK.

Fighting the wrong battle

Scientists who scoff at religious belief miss the point and damage their cause.

Geoffrey Cantor

When modern science began its rise in the seventeenth century, most of the key figures were convinced that its advance would greatly assist religion. For Bacon, Kepler, Boyle, Newton and many others, knowledge of the physical Universe illuminated God's creation. Therefore, through the expansion of science — knowledge of God's works — they expected humankind to come closer to God.

It is ironic that their convictions now appear so misplaced. Although their arguments continued to attract supporters in later centuries, the progress of science has, in the eyes of many, played an active role in the decline of religion. Indeed, some scientists not only assert that science and religion are totally incompatible, but claim that science was responsible for the decline of religion over the past century or two. But this claim is not supported by the evidence. Historians who have investigated why religion — more specifically, mainstream Christianity — appears to have declined in Britain point to an array of factors which have more to do with social, economic and technological changes than with either science in general or any specific scientific theory.

It has become fashionable among scientists with a high media profile to portray religion as the necessary foe of science. This is surely an unwise strategy, one that seems calculated to make scientists appear unreasonable and dogmatic. The tirades of Lewis Wolpert, Richard Dawkins and Peter Atkins also show them to have a very superficial understanding of religion — they would rightly be horrified if others displayed an equal ignorance of science — and not to have noticed that we live in a multi-faith society; that, along with atheists and agnostics, our world is populated by Buddhists, Hindus, Muslims, Jews and Christians of widely different colours, and the many people who possess religious sensibilities but do not subscribe to any organized religion. To attack soft targets in Christianity does not provide an adequate refutation of all forms of religion.

It is true that certain religious groups have been highly critical of science and impeded its advance. We have all heard reports of Southern Baptist preachers inveighing against evolution. Likewise, towards the end of the nineteenth century Roman Catholicism took an anti-science stance. Yet there is another side to the coin — religion has often provided the motivation for pursuing science. Newton and Faraday were two of the many eminent



A confusion of tongues? A babel of arguments obstructs progress across the science–religion divide.

scientists who turned to science to better understand God. They saw no conflict between God's two books — Nature and Revelation. Likewise the vicar–naturalist was a stock character in the eighteenth and nineteenth centuries, labouring on the local geology, flora and fauna for six days a week and treating his congregation to reflections on God's handiwork in his Sabbath sermon.

Religion has also traditionally provided people with communities, with social values and with emotional warmth — aspects of human experience that science cannot offer. Our publicity-seeking scientific clerisy would appear to want to remove these supports and offer nothing in return. Claiming to represent science through the Public Understanding of Science (PUS) initiative, these latter-day Huxleys do science a disservice. These advocates of PUS need to rethink their mission. Taking cheap and uninformed swipes at religion is hardly the best strategy to adopt when trying to encourage

people to take science seriously and become better informed about its methods and content. Likewise, they should not assume that the public at large should revere science and embrace it enthusiastically.

Those who articulate the conflict between science and religion have set the terms of engagement and have forced many religious people into adopting questionable ways of integrating the two domains. Thus we find religious scientists undergoing contortions trying to bridge science and religion through concepts such as indeterminacy in quantum theory. Whatever their validity, such intellectualized responses also fail to tackle many of the most important topics at the science–religion interface, such as the ways in which the values of different faiths lead their members to understand Western science, technology and medicine or, more specifically, how they respond to both physical and mental illness.

Issues of science and religion are important to our civilization — far too significant to be left to either the devoutly religious physicist or the scoffing atheistical biologist. People holding different beliefs and forms of expertise need to work together in an open, non-confrontational environment accepting both science and religion as valid aspects of human experience. It is a challenge facing the coming millennium. ■

Geoffrey Cantor is in the School of Philosophy, University of Leeds, Leeds LS2 9JT, UK.

Newton and Faraday saw no conflict between God's two books — Nature and Revelation.

Avatars in space

Why bother to colonize space when electronic devices do it so much better?

Geoffrey A. Landis

The twenty-first century completed the era when biological species expanded (briefly) their range of habitation into space. Space is a harsh environment for fragile biological machines, which must stay inside carefully climate-controlled cans of air. As we approach the end of the third millennium, our electronic descendants have no such problem. After all, robots are stronger and tougher, don't need life-support systems, and don't get bored. They think of space as home.

Our electronic avatars that thrive in space are about the size of a child's fist (many of them are considerably smaller), and look more like a spider than a spacecraft. Thanks to our avatars, we have spread out across all the planets of our Solar System and many others, and the dark spaces in between as well — but there aren't any humans in space at all, only our electronic ghosts, smarter and faster, engineered for the harsh conditions of space. And maybe every now and then, they think about us biological humans a little, and sometimes even reminisce about living on a planet.

On the surface of the Earth, the superfast, microscopic computers have their applications, too. No task that requires intelligence is left to biological humans. With computer memories that can store information at a density of several bits per atom, an entire biological human's neural state can be encoded in just a few exabytes — less than a milligram of material. When you can scan your entire brain, download it to a chip, and have the chip interact for you inside the virtual worlds of the worldwide interconnected computer network — and then download itself back to you — some people have problems remembering which one 'they' really are. And so humans went into space after all, using high-bandwidth laser links to upload their neural software to space computers.

The space robots, however, continued to evolve and eventually speciate. Some, based on high-temperature silicon-carbide and diamond-based semiconductor materials, moved inward towards Mercury. With copious energy in the form of sunlight, and with resources to be mined from the crust and the polar caps, Mercury seemed the perfect place for machine colonies.

Other machines diverged away from semiconductors, and began to use Josephson junctions for computation. These superconducting machines moved away from the Sun, toward the outermost gas giants and the

Kuiper belt. With superconducting power buses and Josephson electronics, they didn't really need much power.

The main belt asteroids, in between the two realms, served as raw materials for both. The divergence of electronic species led to inevitable conflict over resources that were valuable to both. Even now, wars are being fought between the machines over resources in this middle area; no doubt fiercer wars are yet to come, as ever larger and more ambitious projects make resources more valuable.

The planets themselves — outside Mercury — are of only minor interest to our electronic avatars. With too much free oxygen and liquid water, and unreliable sunlight, Earth is a poor place for machines.

The avatars have their own society. With several trillion individuals, their society is so complicated that it will never be possible for a biological human to understand it. Their art, their music, their intricate social structure — the humans that have uploaded into space are only a negligibly small part of it.

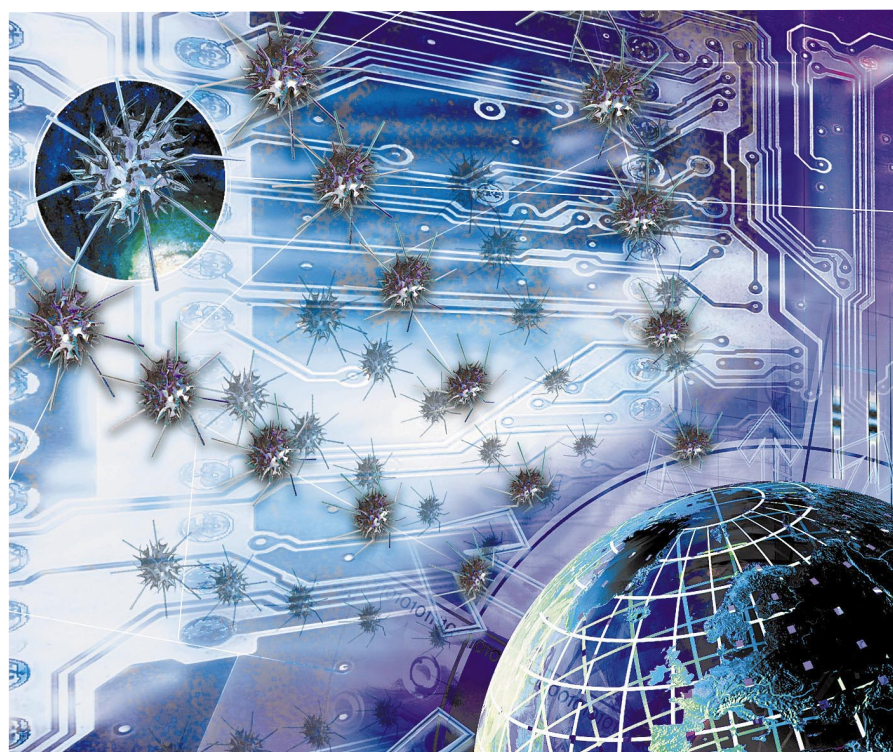
Far below, and almost ignored by the enormous electronic ecosystem in space, biological species still thrive on Earth. No longer on the cutting edge of science — every problem that could possibly be understood by a biological mind has long ago been studied to completion by silicon-carbide brains — the humans devote themselves to the art of living

graciously, in harmony with an ecosystem that is now understood in exhaustive detail. Art, history, literature, sports, recreation, religion, philosophy: humans still have plenty of activities to which they can devote themselves. Work and war are equally forgotten — what would be the point? Those who want territory, or resources, or power, eventually decide to upload themselves into space.

As the millennium draws to a close, the Josephson computers are now reaching for the stars the slow way, colonizing each one of ten trillion comets in the dark spaces between the stars. The Mercury machines, on the other hand, are reaching for the stars in the fast lane: they have constructed an enormous particle-beam accelerator, closer to the Sun than Mercury. Machines no larger than a grain of rice are sailing on particle beams to 500 of the nearest stars, at 89% of the speed of light. These machines are only there to fly past and report, but if they find conditions hospitable — and for machines that have no need for planets, the range of 'hospitable' is very wide indeed — the next generation will be colonists.

The future looks interesting.

Geoffrey A. Landis was a member of the Mars Pathfinder science team, and is currently working on hardware for future Mars missions. His first novel, *Mars Crossing*, will appear in late 2000. <http://www.sff.net/people/geoffrey.landis>



JACEY

Proteomics for the pore

Günter Blobel and Richard W. Wozniak

The nuclear pore complex is the gateway that regulates two-way traffic between the nucleus and the rest of the cell. The architecture of this huge structure is now revealed in unprecedented detail.

Rarely in the field of cell biology do we catch a glimpse of the complete molecular picture of a major subcellular structure. Until recently, the complexity of such structures presented an analytical challenge that seemed almost insurmountable. Now, through the use of revolutionary new technologies, even small teams of researchers can catalogue the components of large cellular machines, map their architecture, and finally begin to grasp how they work. This is beautifully demonstrated by an article, published earlier this week in the *Journal of Cell Biology*, in which Rout and colleagues use proteomics to reveal the complete molecular architecture of the yeast nuclear pore complex¹. This work will point the way towards unveiling the secrets of other giant molecular machines, such as the centrosome, a player in the process of cell division.

Nuclear pore complexes are massive octagonal structures that extend across a cell's nuclear envelope (Fig. 1), forming gateways that govern the flow of macromolecules (carried as cargo by soluble transport factors) between the cytoplasm and the nucleus. Here they can influence a vast array of signal-transduction and gene-expression pathways. In recent years, a variety of approaches has been used to define the components of the complex², and with considerable success. But it was unknown to what degree the emerging inventory represented its full complement of proteins. Now, using state-of-the-art mass-spectrometry techniques, Rout and co-workers¹ have identified all of the detectable polypeptides in complete yeast nuclear pore complexes³, and have systematically localized every potential component protein within the cell. Surprisingly, only around 30 different proteins proved to be components of the complex, a number that is considerably lower than previous estimates. After all, the ribosome, the cell's protein-synthesizing factory which is one-twentieth the mass of the nuclear pore complex, contains 75 different proteins.

How could this relatively streamlined set of components account for the massive structure of the nuclear pore complex? To address this question, Rout *et al.* localized each protein within the complex by immunoelectron microscopy, and determined how much of each was present. Together with morphological data on the yeast nuclear pore complex obtained by electron

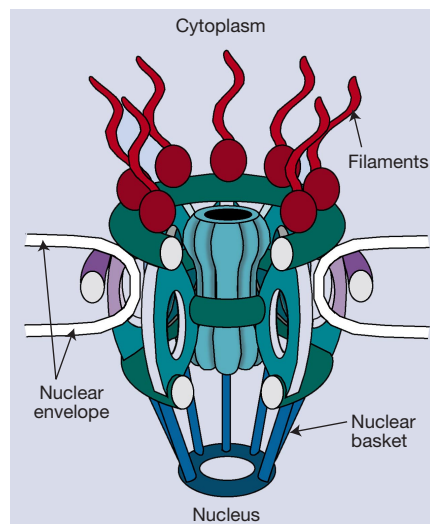


Figure 1 The nuclear pore complex. Rout *et al.*¹ show that it consists of only about 30 different proteins, forming a basic subunit that is repeated 16 times. There are additional, filamentous structures on both sides of the pore. On the nuclear side, these link together to form the nuclear basket.

cryomicroscopy⁴, their results indicate that the complex is primarily composed of a single substructure repeated 16 times, with eightfold radial symmetry on the cytoplasmic face and a corresponding mirror image on the side that faces into the nucleus. Interestingly, this substructure includes many proteins that have been previously implicated as binding sites for the cargo-carrying transport factors. Built upon this symmetrical core are additional transport-factor-binding proteins, which contribute to filamentous structures on the cytoplasmic and nuclear sides of the pore (Fig. 2). Based on measurements of the abundance of these proteins, the authors estimate that there are more than a hundred such binding sites on the surface of the nuclear pore complex.

Conspicuously absent from their list of components are any obvious motor proteins (such as myosin), which have long been speculated to be missing players in the transport machinery. As transport between the nucleus and the cytoplasm can proceed against a concentration gradient of cargo, energy must be supplied to the system. Many groups have shown that the GTPase enzyme Ran is a prime candidate for providing the necessary energy, apparently being main-

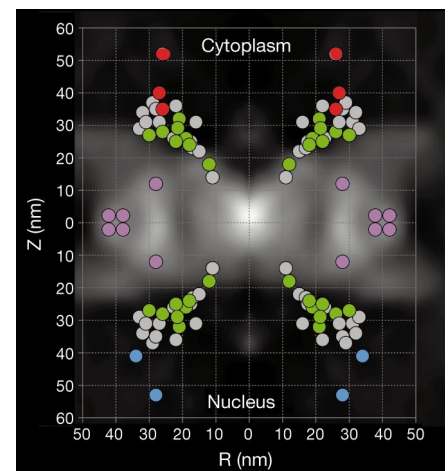


Figure 2 The mapped positions of each of the proteins in the yeast nuclear pore complex. The positions are shown relative to the central axis (R) and a plane passing through the centre of the complex and parallel to the nuclear envelope (Z), and overlies a cross-section of the yeast nuclear pore complex (from refs 1 and 4). Green and grey circles represent symmetrically located proteins with positions mirrored on both faces of the complex. The approximate positions of the asymmetric proteins are shown in red and blue. The pink circles denote membrane proteins that associate with the nuclear pore complex. The specific protein represented by each circle is defined in Rout *et al.*¹.

tained in the GTP-bound form within the nucleoplasm and the GDP-bound form in the cytoplasm.

With these considerations in mind, Rout *et al.* present a model for how the nuclear pore complex mediates selective transport across the nuclear envelope. They propose that transport depends on several factors: diffusion; affinity of transport factors for the nuclear pore complex; differences in these affinities between symmetric and asymmetric binding sites; and the distribution of Ran-GTP and Ran-GDP. The complex is known to present a considerable barrier to the diffusion of macromolecules. However, cargo-laden transport factors destined to pass through the pore have an advantage — they bind to the complex. Because this results in an increase in local concentration around the central channel, they are more likely to pass through the diffusion barrier of the pore than proteins that do not bind. In this model, the nuclear pore complex behaves

like a 'virtual gate': selectivity is achieved without necessarily invoking large concerted movements (as, for example, in the opening and closing of a shutter). This is then followed by trapping at high-affinity binding sites on the opposite face of the complex. Finally, the transport reaction is terminated by exposure to the alternative Ran environment to that from which the transport factor started. In this way, Ran GTPase injects energy into the transport cycle.

Although the model is consistent with data from many groups, and suggests further interesting experiments, it has yet to be tested directly. Furthermore, the model is certainly a simplified one. For example, some transport pathways may not use Ran⁵. Moreover, the situation appears to be more complex in multicellular organisms, as their asymmetric structures appear to be laden with binding sites for Ran and other Ran-regulatory proteins.

Not only does the work of Rout and his collaborators provide us with a vast amount of information on the molecular organization of the nuclear pore complex, it also gives

us an inkling of the future of cell biology. Surely, similar approaches will soon lead to the comprehensive characterization of other macromolecular machines and organelles^{6–8}. As this new information accumulates, we will be confronted with the tasks of understanding the dynamics of protein–protein interactions and tackling molecular mechanisms. Cell biologists can now enjoy the riches of our post-genome era. ■

Günter Blobel is in the Laboratory of Cell Biology, Howard Hughes Medical Institute, Rockefeller University, 1230 York Avenue, New York, New York 10021-6399, USA.

e-mail: blobel@rockvax.rockefeller.edu

Richard W. Wozniak is in the Department of Cell Biology, 514 Medical Sciences Building, University of Alberta, Edmonton, Alberta, T6G 2H7, Canada.

e-mail: rick.wozniak@ualberta.ca

1. Rout, M. P. *et al.* *J. Cell Biol.* **148**, 635–651 (2000).
2. Fabre, E. & Hurt, E. *Annu. Rev. Genet.* **31**, 277–313 (1997).
3. Rout, M. P. & Blobel, G. *J. Cell Biol.* **123**, 771–783 (1993).
4. Yang, Q., Rout, M. P. & Akey, C. W. *Mol. Cell* **1**, 223–234 (1998).
5. Hetzer, M. & Mattaj, J. W. *J. Cell Biol.* **148**, 293–304 (2000).
6. Wigge, P. A. *et al.* *J. Cell Biol.* **141**, 967–977 (1998).
7. Neubauer, G. *et al.* *Nature Genet.* **20**, 46–50 (1998).
8. Ogryzko, V. V. *et al.* *Cell* **94**, 35–44 (1998).

Crystal structure

As weird as they come

Volker Heine

Most metals have simple structures, such as face-centred cubic, which can be obtained by packing marbles into a box. But there are exceptions, notably gallium's unique structure in which each atom has seven near neighbours, and the distorted structures of mercury and tin. These are the only metals with unusual structures at normal pressures and temperatures, but there are other examples at high pressures¹. Richard Nelmes and collaborators² have now reported in *Physical Review Letters* what must be the weirdest known atomic structure of a metal, indeed of any pure element. Using X-ray diffraction they determined the crystal structure of the metal barium under pressures from 12 to 45 gigapascals (120,000 to 450,000 atmospheres). At high pressure most materials change the atomic arrangements that make up their crystal structures, with barium showing five structures in all, of which the latest one is number IV.

Nelmes *et al.*² describe the structure of barium IV as having two parts with different periodicity — a 'guest' structure stacked inside tubular holes of a 'host' structure (Fig. 1a). It is like having two ladders side by side with different step heights. If the step height of one ladder is, say, 43/49 of the other, then one would have to go 49 rungs of the first ladder (and 43 rungs on the other) for the rungs to be at the same level. If the ratio of the step

heights is an irrational number, like π or the square root of two, then the structure is said to be 'incommensurate' because there is no distance at which it repeats exactly. There are many materials that have incommensurate structures, such as organic inclusion compounds where the host and guest molecules are chemically different.

What is surprising about barium IV is

that the guest and host are made from the same pure element. The host structure consists of hollow octagonal cylinders of barium atoms, inside which is the guest structure — straight chains of yet more barium atoms (Fig. 1a). Nelmes *et al.* find that the repeat distance along the guest chains ($c_g = 3.4095$ ångström, where $1 \text{ Å} = 0.1 \text{ nm}$) is quite independent of the repeat distance along the host cylinders ($c_h = 4.6966 \text{ Å}$). Moreover, the ratio c_g/c_h appears to be irrational and varies smoothly with pressure, so the chains and guest cylinders do not fit well together at any of the pressures studied (Fig. 1b).

How can we make sense of such a structure? The unusual structure of gallium, where each atom has seven near neighbours, comes about from a competition between two different types of force, with the structure trying to serve both masters^{3,4}. First, there is the interatomic bonding that depends on the separation between atoms, and which determines the preferred distance for an atom's near neighbours. In metals there is a second force that depends on the volume available per atom. This force is due to the quantum mechanics of the conducting electrons being free to move through the metal as a 'nearly free electron gas'. The smaller the volume available to the electrons, the shorter is their quantum-mechanical wavelength and the higher their energy. The combination of these two forces determines how many near neighbours an atom will have, which in turn restricts the crystal structure^{3,4}.

In barium IV, each host atom has about nine near neighbours at a distance of about 3.4 Å, whereas each guest atom has about ten such neighbours. There are no simple structures with nine to ten near neighbours — this arrangement requires a complex structure, such as the one observed. Calculations of the distance and volume forces should confirm this general picture.

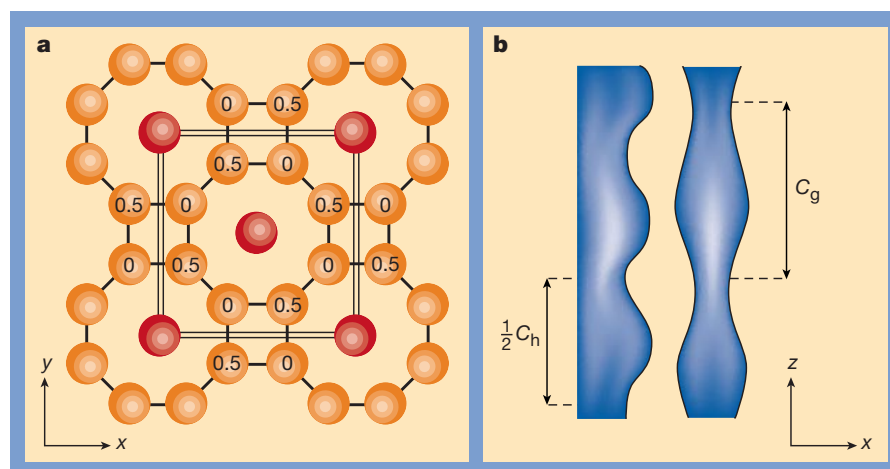


Figure 1 The high-pressure structure of barium IV discovered by Nelmes *et al.*². a, Structure projected onto the x–y plane, showing the atoms of the host structure (yellow) with heights in units of c_h and the guest chains (red). b, Electron density map in the x–z plane of the atomic chain (right) with repeat distance c_g and the side of a host cylinder (left). The ratio c_g/c_h is not a rational number, making the structure incommensurate — that is, there is no distance at which it repeats exactly.

There is another quantum-mechanical wrinkle that may need investigating. Electron waves have a complicated shape as they move through a metal, but they can still be described by a three-dimensional wave vector, k , with energy $E(k)$. The quantum states, k , are filled with electrons from the lowest energy up, just as in the shell structure of an atom. In a metal, electrons run out at some energy E_F , called the Fermi energy, such that all states below the Fermi surface are filled up. The standing wave formed by electrons at wave vectors k and $-k$ at opposite ends of the Fermi surface introduces a non-uniform electron density with repeat distance $2\pi/(2k)$, which in general is incommensurate with the underlying crystal structure. Further calculations are needed to assess the importance of this effect in barium IV, but it has been seen clearly in chromium and some conducting compounds.

Finally, does nature really like incommensurability or will the host and guest structures always expand or contract a little to make c_g/c_h a rational number? That way the distances between all the host and guest atoms could be to some extent optimized, whereas in a truly incommensurate structure none of them ever fits perfectly. For these

reasons, scientists have often felt sceptical about the existence of a truly incommensurate structure, particularly at a temperature of absolute zero. Experimentally it is difficult to distinguish between an incommensurate structure with an irrational repeat, such as $\sqrt{2}$, and one with a rational repeat, such as 1.4142. However, this issue was resolved theoretically using a precise mathematical model⁵. The answer is that if the atoms of the chain can slide up the host cylinders fairly smoothly, then the structure is truly incommensurate. But if the chain–host interaction (Fig. 1b) were bumpier than a certain critical value, then the structure would lock on and jump in steps from one rational value of c_g/c_h to another, as the pressure is varied. We do not yet have a full theory of the structure of barium under pressure found by Nemes *et al.*, but we have the intellectual building bricks that are needed. ■

Volker Heine is at the Cavendish Laboratory, Madingley Road, Cambridge CB3 0HE, UK.
e-mail: vh200@phy.cam.ac.uk

1. Soderlind, P. *et al.* *Nature* **374**, 524–525 (1995).
2. Nemes, R. J., Allan, D. R., McMahon, M. I. & Belmonte, S. A. *Phys. Rev. Lett.* **83**, 4081–4084 (1999).
3. Heine, V. & Weaire, D. *Solid State Phys.* **24**, 249–463 (1970).
4. Hafner, J. & Heine, V. *J. Phys. F* **13**, 2479–2501 (1983).
5. Peyrard, M. & Aubry, S. *J. Phys. C* **16**, 1593–1608 (1983).

Neurobiology

Tepid tastes

Robert A. Frank

When we eat, foods activate not only taste but also smell, thermal, tactile and irritant receptors. How do these perceptions interact to create the sensation of flavour? As the dimensions of flavour are associated with independent sensory systems, the sensory interactions that accompany food consumption are thought to occur at higher, cognitive levels of the brain. But persistent hints of a relationship between taste perception and thermal stimulation of the tongue at the early stages of gustatory coding have been reported.

On page 889 of this issue, Cruz and Green¹ show for the first time that changing the temperature of a small area of the tongue can create the sensation of taste in humans. This raises questions about the transduction and coding of taste signals, and may reflect a unique interaction between the gustatory and somatosensory (touch- and temperature-sensing) systems.

Cruz and Green controlled the temperature of areas measuring about 1 cm² on the tip or the back of the tongue. The area was either warmed from 20 to 35 °C, or cooled from 35 °C to either 15 or 5 °C. (The hottest and coolest temperatures are well within the range of temperatures normally produced by

hot foods or ice-cream; Fig. 1.) For some people, warming the front of the tongue produced a weak but clear sweet sensation, whereas cooling resulted in sour or salty tastes. Warming the back of the tongue produced only very weak experiences of sweetness, but cooling produced moderate bitter or sour sensations. Some people reported all



Figure 1 Cool taste. Cruz and Green¹ show that changing the temperature of small areas of the tongue's surface can cause the sensation of taste — but these girls' enjoyment of their ice-cream probably stems from the flavours in the ice-cream itself.

of these effects, others a few, and others none of them. These individual differences are reminiscent of other differences in taste sensitivity, such as taste 'blindness' to thiourea compounds².

Thermal and gustatory receptors are close together in the mouth³. Nerve fibres from two cranial nerves innervate the papillae — tiny, bulb-like structures on the surface of the tongue (Fig. 2). Fibres from the chorda tympani branch of the facial nerve terminate in taste buds on the papillae. These fibres detect sweet, salty, sour and bitter tastes⁴. Fibres from the lingual branch of the trigeminal nerve terminate in the epithelium of the papillae, close to taste buds, and sense temperature, pain, irritation and pressure⁵.

Until now, there was only a little, indirect evidence that non-gustatory stimulation alone could produce tastes. Georg von Békésy⁶ proposed a controversial theory of taste coding that involved the interaction of temperature and taste. Based on studies showing an interaction between taste and thermal stimuli applied on the two sides of the tongue, von Békésy suggested two classes of stimuli: warming/sweet/bitter and cooling/salty/sour. Cruz and Green speculate that von Békésy may have observed the fusion of chemical and thermally induced tastes across the tongue. Many gustatory neurons respond to temperature changes⁷, but it has been assumed that these responses represent noise that is filtered out in the central nervous system.

The simplest explanation for thermal taste is that changes in temperature somehow activate the receptor mechanisms responsible for normal gustatory coding. If this is correct, interfering with normal taste might block thermal taste. Alternatively, thermal taste might be mediated through taste cells, but not by the normal receptor mechanism, in which case blocking taste receptors would have no effect. These competing hypotheses can be tested by studying the effects of taste blockers on thermal taste, as these substances act early in transduction (for example, extracts of the plant *Gymnema sylvestre* eliminate sweet tastes⁸).

Another possibility is that thermal taste is mediated by the somatosensory system. Lingual nerve fibres terminate close to taste cells and carry information to areas of the brain primarily involved in gustatory processing, and many of these neurons respond to salt and sugar solutions⁹. Might these fibres be involved in thermal taste?

People do not normally report thermal tastes, probably because we do not usually experience the stimulation conditions used by Cruz and Green. We do not restrict foods to a single location on the tongue, nor do we routinely consume tasteless hot or cold foods. Thermal tastes appear to be weak compared with the tastes of foods and beverages, so they are probably masked under

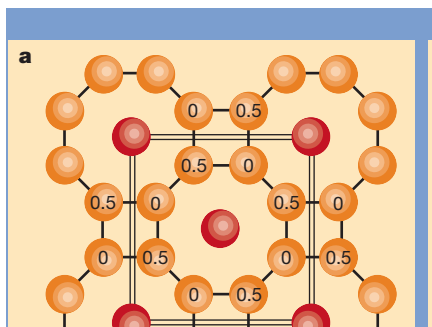


Figure 2 Innervation of the papillae by the gustatory and somatosensory systems. Taste information is coded by specialized sensory cells in taste buds, whereas the nerve endings of the lingual nerve mediate sensations of temperature, touch, irritation and pain. The thermal tastes described by Cruz and Green¹ may be mediated by either of these pathways.

normal circumstances. (You can, however, discover whether you are sensitive to thermal saltiness, the rarest form of thermal taste in Cruz and Green's study, by holding an ice cube to the tip of your tongue.)

Although its everyday function may be limited, thermal taste illustrates the complexities of sensory coding. From that per-

spective, the problem is not so much why thermal tastes occur, but rather, how they are avoided. The taste system extracts specific gustatory information from receptor cells responding broadly to changes in their physical and chemical environment. How is the taste signal extracted from a noisy background induced by fluctuations in temperature and other non-gustatory stimuli? Understanding thermal taste might give us insight into how the nervous system achieves this end.

Robert A. Frank is in the Department of Psychology, University of Cincinnati, Mail Location 627, Cincinnati, Ohio 45221-0627, USA.

e-mail: robert.frank@uc.edu

1. Cruz, A. & Green, B. G. *Nature* **403**, 889–892 (2000).
2. Hall, M. J., Bartoshuk, L. M., Cain, W. S. & Stevens, J. C. *Nature* **253**, 442–443 (1975).
3. Whitehead, M. C., Ganchrow, J. R., Ganchrow, D. & Yao, B. *Neuroscience* **93**, 931–941 (1999).
4. Dodd, J. & Castellucci, V. F. in *Principles of Neuroscience* (eds Kandel, E. R., Schwartz, T. M. & Jessell, T. M.) 512–529 (Elsevier, New York, 1991).
5. Dodd, J. & Kelly, J. P. in *Principles of Neuroscience* (eds Kandel, E. R., Schwartz, T. M. & Jessell, T. M.) 701–710 (Elsevier, New York, 1991).
6. von Békésy, G. *Science* **145**, 834–835 (1964).
7. Oakley, B. *Chem. Senses* **10**, 469–481 (1985).
8. Frank, R. A., Kennedy, L. M., Mize, S. J. S., de los Santos, H. C. & Green, S. *Chem. Senses* **17**, 461–479 (1992).
9. Pittman, D. W. & Contreas, R. J. *Brain Res.* **790**, 224–235 (1998).

Earth science

A shear pathway to the core

Michael J. Walter

A popular model for the formation of the Earth's core has iron-rich liquid metal segregating gravitationally from a molten silicate mantle (a so-called magma ocean). There are three lines of support for this model, one of which is based on the physics of metal segregation. Experiments have shown that iron-rich metallic melt cannot segregate effectively from a solid silicate mantle¹, which is an alternative possibility, and so have indicated that mantle melting — a magma ocean — was required for metal segregation to the core.

On page 883 of this issue, Bruhn *et al.*² seriously undermine this evidence. They show that, in a solid silicate mantle that is deforming under the action of shear stresses, iron-rich metallic melt may form interconnected channels that would allow effective drainage of metal to the core. The key factor here is the inclusion of shear forces in their experiments.

Apart from the physical evidence for extensive melting of the proto-Earth, theoretical and geochemical evidence are also used to prop up the magma-ocean model. Theory indicates that most of the Earth's mass stems from the impact of large objects with the proto-Earth, and that these energetic

collisions would have melted a large fraction of the silicate mantle^{3,4}. Geochemistry tells us that the mantle abundance of several siderophile (iron-loving) elements, notably nickel and cobalt, may have been set by equilibrium between liquid metal and silicate melt at high pressures and temperatures⁵. Neither of these two lines of evidence is robust enough to provide conclusive support for a magma ocean, so the addition of the physical evidence has made the magma-ocean model more credible.

Earth's upper mantle is composed mostly of olivine, a magnesium–iron silicate mineral. Under static conditions — that is, in the absence of any shear stresses — the ability of metallic melt to separate and move through the solid upper mantle depends primarily on the relative surface energies of the melt–solid and solid–solid interfaces. If the melt–solid interface has a relatively larger surface energy, then melt will tend to form pockets and become trapped at grain-boundary triple junctions, at least until a high melt fraction is reached. Conversely, a relatively smaller melt–solid surface energy promotes dissemination of melt along grain boundaries, often referred to as 'wetting', so melt becomes easily inter-

connected and permeability is established.

Melt interconnectedness is a prerequisite for density differences between melt and solid to drive melt away from its source by percolation. For example, silicate melt in the upper mantle readily wets the grain boundaries of silicate minerals and, because of its relatively low density, the melt moves upwards to the surface, erupting as lava from a volcano. Similarly, if relatively high-density metallic melt could wet silicate grain boundaries, it would drain downwards to the Earth's core.

Experiments carried out under nominally static conditions have shown that the surface energetics between iron-rich metallic melt and olivine are very unfavourable for wetting grain boundaries, inhibiting metal segregation from the upper mantle and percolation to the core^{1,6}. However, if the silicate minerals are themselves substantially melted then the silicate melt can provide the pathways required for metal segregation. Even more simply, metal would easily sink to the core if the mantle were wholly melted. So the logic has been as follows: in the absence of silicate melting, iron-rich metallic melt would not be able to segregate effectively and drain from the upper mantle; but in a partly molten mantle, or in a magma ocean, metal could easily segregate to form the core. The proverbial fly in the ointment is that the solid upper mantle is anything but static.

Heat is taken from the centre to the surface of the Earth primarily by convection in the mantle. For the past four billion years or so, the mantle has been solid, with the exception of relatively shallow melting regions. Even so, over this long timescale the mantle behaves like boiling water on a stove. Hot, low-density material from deep within the Earth moves up towards the surface where it loses heat by conduction, becomes denser, and sinks back down towards the bottom. The differential buoyancy between the hot and cold mantle material creates shear stresses in the mantle. Convective motion occurs by plastic deformation of the minerals in the presence of such stresses. Presumably, the proto-Earth was hotter than the Earth is now, so its mantle may have been subject to much greater deforming forces. At any rate, shear stresses in the mantle are the rule, not the exception.

Bruhn *et al.*² have performed experiments at high pressures and temperatures in which mixtures of metal and olivine were subjected to high shear stresses. They discovered that, because of the deformation of the mineral grains, the initially trapped pockets of metallic melt become elongated and aligned along a preferred orientation relative to the principal shear direction. Three-dimensional analysis of the experimental samples indicates that the melt pockets are interconnected, thereby providing the channels necessary for percolation. This result is a



100 YEARS AGO

Prof. E. H. Barbour, professor of geology in the University of Nebraska, has recently given reasons for believing that a rapid decline of geyser activity is taking place in the Yellowstone National Park. If the present rate of decline continues, it seems possible that within a decade many of the well-known geysers will have died out. As a result of an examination of the geyser area, after an interval of four years, Prof. Barbour gives the following instances among others of the diminution of activity which has occurred: The fountain Geyser, which was such a favourite that the Fountain Hotel was situated at that spot, is now wholly extinct, and a very inferior substitute named the Dewey Geyser has taken its place. The Cascade Geyser, another favourite because of the frequency of its eruptions (about every fifteen minutes), has dropped to an eruption interval of once every twenty-four hours. The Grand Geyser, which used to burst out once a day, was only active three or four times the past season. The Beehive Geyser, active in 1895, is supposed to be wholly extinct. Old Faithful seems as fine as ever, but the interval of eruption is now about seventy-five or eighty minutes instead of once an hour. If it is possible to judge fairly of such matters, there seems to be an increase in the ebullition of the water in that greatest of geysers, the Excelsior, which leads to a feeble hope that it may possibly be rejuvenated yet once again. An apparent increase in the activity of the Mud Geyser has also been remarked; but in spite of these cases, on the whole it appears that a distinct decline of activity is taking place.

From *Nature* 22 February 1900.

50 YEARS AGO

During two and a half decades of African bush geology, an observation has forced itself upon me, which may perhaps have been noticed by others, namely, that elephants appear to show preference for certain geological horizons. This is at times so well pronounced that the boundary line between two formations can be deduced from native information as to how far elephants circulate in the district. I have noticed it particularly in regions where crystalline rocks, mainly granite masses, come into contact with sandstone, and it is the latter which the elephants prefer.

From *Nature* 25 February 1950.

serious challenge to the conventional view, based on the results of static experiments, that iron-rich metallic melt cannot separate effectively from an olivine-rich solid mantle.

The implication for models of core formation is that metal could have segregated in the proto-Earth and other terrestrial bodies with or without silicate melting. So one leg of the tripod of support for the core-segregation model involving a magma ocean may have just been sheared away. However, this result in no way implies that the Earth's core did not segregate from a magma ocean. Rather, it means that proponents of the magma-ocean model can no longer invoke

the argument that melting of the upper mantle was required for metal to segregate and drain to the core.

Michael J. Walter is at the Institute for Study of the Earth's Interior, Okayama University, Misasa, Tottori 682-0193, Japan.

e-mail: walter@misasa.okayama-u.ac.jp

1. Minarik, W. G., Ryerson, F. J. & Watson, E. B. *Science* **272**, 530–533 (1985).
2. Bruhn, D., Groebner, N. & Kohlstedt, D. L. *Nature* **403**, 883–886 (2000).
3. Wetherill, G. W. *Science* **228**, 877–879 (1985).
4. Benz, W. & Cameron, A. G. W. in *Origin of the Earth* (eds Newsom, H. E. & Jones, J. H.) 61–68 (Oxford Univ. Press, New York, 1990).
5. Li, J. & Agee, C. *Nature* **381**, 686–689 (1996).
6. Shannon, M. C. & Agee, C. B. *Geophys. Res. Lett.* **23**, 2717–2720 (1996).

Plant biology

A social stigma

Teh-hui Kao and Andrew G. McCubbin

In many bisexual flowering plants, the close proximity of male and female reproductive organs has resulted in the development of strategies to prevent inbreeding. One strategy, called self-incompatibility, enables the pistil (female) tissue to distinguish between self (genetically related) and non-self (genetically unrelated) pollen. How is this achieved?

Species of *Brassica* (cabbage) have been widely used in addressing this question. On page 913 of this issue, Takasaki *et al.*¹ show that a gene of *Brassica campestris*, named *SRK* (S-locus receptor kinase), controls the pistil's ability to recognize and reject self pollen. This finding, coupled with identification of the gene that controls pollen function in self-incompatibility^{2,3}, and of a potential substrate of SRK protein⁴, has brought us closer to understanding the molecular and biochemical bases of self-incompatibility in *Brassica*.

The S locus is polymorphic and confers genetic identity (termed S haplotype specificity) on the pollen and pistil of self-incompatible plants; because of their inability to self-fertilize, these plants are heterozygous with respect to the locus. The outcome of pollination depends on the S haplotypes of male and female parents. In the simplest case, if either of the two S haplotypes of the male parent matches one of those of the female parent, pollen is recognized as 'self' and rejected.

The S locus gene or genes determining the S haplotype specificity of the pistil would be expected to show S-specific sequence polymorphism and be expressed in the stigma (on whose surface pollen rejection occurs). Two such genes have been identified: *SLG* (S locus glycoprotein)⁵ and *SRK*⁶, and it has been proposed that both are required for self-incompatibility⁶. But previ-

ous attempts to obtain direct evidence for their function were inconclusive. Plants transformed with either *SLG* or *SRK* of a new S haplotype were rendered self-compatible. This was due to a phenomenon termed homology-dependent gene silencing, in which the expression of the transgene, as well as endogenous *SLG* and *SRK*, were all suppressed⁷. Moreover, although suppression of endogenous *SRK* or *SLG* by antisense *SRK* or *SLG* resulted in self-compatibility, it simultaneously reduced transcript levels of endogenous *SRK* and *SLG*, as well as of other *SLG*-related genes. So it was difficult to assess the function of each gene⁸.

Takasaki *et al.*¹ overcame the problem of homology-dependent silencing by introducing *SRK* and *SLG* of S²⁸ haplotype into plants that carried S⁶⁰ and S⁵²/S⁶⁰ haplotypes, respectively; the *SLGs* and *SRKs* of these haplotypes have a low degree of sequence similarity with *SLG*²⁸ and *SRK*²⁸ transgenes. Self-incompatible transgenic plants expressing the *SRK*²⁸ transgene were obtained, and Takasaki *et al.* found that their pistils, but not pollen, had acquired the S²⁸ specificity. In contrast, the pistils of a transgenic plant expressing a normal level of the *SLG*²⁸ transgene failed to reject pollen of S²⁸ haplotype. So for the first time we have direct evidence that *SRK*, but not *SLG*, determines S haplotype specificity in the pistil.

The SRK protein has structural and biochemical properties characteristic of receptor kinases: an extracellular domain (called the S domain because of sequence similarity to *SLG*); a transmembrane domain; and a cytosolic domain with serine/threonine kinase activity. By analogy with animal receptor kinases, the extracellular domain is thought to interact with a pollen ligand, setting off a cascade of biochemical reactions which results in inhibition of pollen

germination (Fig. 1). The pollen ligand is probably the small, basic cysteine-rich protein (named SCR or SP11) that is the pollen S haplotype determinant^{2,3}. To date, only one of the components of the SRK-mediated signalling pathway has been positively identified⁴. This is a protein named ARC1, which interacts with the kinase domain of SRK *in vitro*, and which, when suppressed in stigmas of transgenic plants, results in self-compatibility. Although the cellular function of ARC1 is unknown, it can serve as a handle for identifying other components of the signalling pathway.

In most cases, the SLG of a given S haplotype has the greatest sequence similarity with the extracellular domain of the SRK of the same S haplotype. This apparent coevolution of SLG and SRK is puzzling if SLG is not involved in S haplotype specificity. Takasaki *et al.*¹ show that, although SRK alone is sufficient for the pistil S haplotype specificity, the self-incompatibility response is strongest if SLG of the same S haplotype is also present.

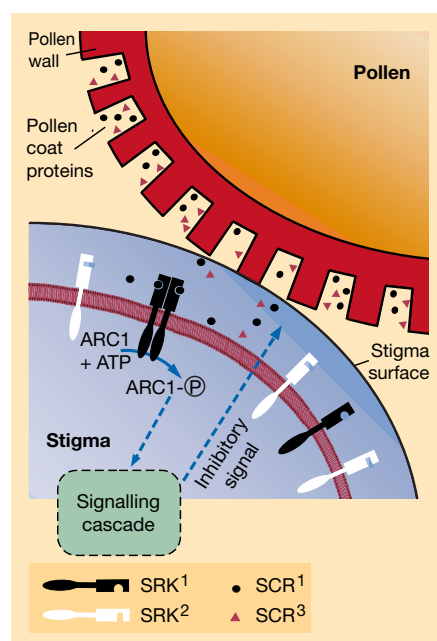


Figure 1 Model for SRK-mediated self-incompatibility in *Brassica*, stemming from the work of Takasaki *et al.*¹. A pollen grain from a plant with *S*¹*S*³ genotype carries two different S haplotype determinants, SCR¹ and SCR³, in the pollen coat. When this pollen grain makes contact with the stigma of a plant with *S*¹*S*² genotype, both pollen determinants are released and taken into the wall of a papillar cell (the epidermal cell of the stigma). There, SCR¹ interacts with the extracellular domain of SRK¹, setting off a cascade of biochemical reactions (of which phosphorylation of ARC1 is the only one known). The end result is inhibition of pollen germination. Although SRK¹ is shown as a dimer, this may not be its active form. SCR¹ is probably (but not certainly) the ligand of SRK¹. Possible interactions between SRK and SLG, as proposed by Takasaki *et al.*, are not shown.

They suggest that SLG facilitates the SRK-mediated response through interaction with the S domain of SRK. So it is possible that coevolution of SLG with SRK is necessary to ensure high levels of self-incompatibility. The transgenic plants analysed by Takasaki *et al.* produced only about 30% of the wild-type level of SRK, however, so it remains possible that SLG is not required when normal levels of SRK are present.

Previously, SLG has been shown to be required for adhesion of pollen to the stigma⁹. This may involve interactions of SLG with pollen-coat proteins belonging to the same family as SCR/SP11 (ref. 10). Although such interactions are not S-haplotype specific, the parallel between these and the presumed SRK/SCR interactions is striking. One could speculate that the *Brassica* self-incompatibility system might have evolved by hijacking a pre-existing system used in the fundamental process of pollination.

Although *Brassica* species are generally self-incompatible, most other agricultural species (tomato, maize and soybean, for instance) are self-compatible. Plants raised from hybrid seed give higher yields than plants raised from self-pollination owing to a phenomenon called hybrid vigour. Self-incompatibility would be an effective method of preventing self-fertilization in the production of hybrid seed. So a better understanding of self-incompatibility systems, such as that in *Brassica*, will bring us closer

to achieving the goal of introducing self-incompatibility into other crop plants to facilitate hybrid seed production.

Clearly, discovering that SLG is not required for S haplotype specificity is as significant as finding that SRK is required. Further challenges include identifying the components downstream of ARC1 in the SRK signalling pathway, the mechanism of pollen rejection, and how the male (*SCR/SP11*) and the female (*SRK*) genes coevolve in the generation of new specificities. The study of self-incompatibility in *Brassica*, if not its action, is beginning to yield fruit.

Teh-hui Kao and Andrew G. McCubbin are in the Department of Biochemistry and Molecular Biology, 403 Althouse Laboratory, Pennsylvania State University, University Park, Pennsylvania 16802-4500, USA.

e-mails: txk3@psu.edu

agm3@psu.edu

1. Takasaki, T. *et al.* *Nature* **403**, 913–916 (2000).
2. Schopfer, C. R., Nasrallah, M. E. & Nasrallah, J. B. *Science* **286**, 1697–1700 (1999).
3. Takayama, S. *et al.* *Proc. Natl Acad. Sci. USA* **97**, 1920–1925 (2000).
4. Stone, S. L., Arnold, M. & Goring D. R. *Science* **286**, 1729–1731 (1999).
5. Nasrallah, J. B., Kao, T.-H., Chen, C.-H., Goldberg, M. L. & Nasrallah, M. E. *Nature* **326**, 617–619 (1987).
6. Stein, J. C., Howlett, B., Boyes, D. C., Nasrallah, M. E. & Nasrallah, J. B. *Proc. Natl Acad. Sci. USA* **88**, 8816–8820 (1991).
7. Conner, J. A. *et al.* *Plant J.* **11**, 809–823 (1997).
8. Shiba, H., Hinata, K., Suzuki, A. & Isogai, A. *Proc. Jpn. Acad.* **71**, 223–225 (1995).
9. Luu, D.-T., Marty-Mazars, D., Trick, M., Dumas, C. & Heizmann, P. *Pl. Cell* **11**, 251–262 (1999).
10. Doughty, J., Hedderson, F., McCubbin, A. & Dickinson, H. G. *Proc. Natl Acad. Sci. USA* **90**, 467–471 (1993).

Astronomy

Galaxy formation in action

Gary Welch

Understanding galaxy formation is one of astronomy's great challenges. Detailed images from the Hubble Space Telescope show that galaxies frequently interacted in the past¹, suggesting that interactions were important in creating the bright galaxies we see today^{2,3}. The work by Braine *et al.*⁴ on page 867 of this issue provides important evidence that interactions continue to form new galaxies from recycled debris. If true, it would offer astronomers the exciting prospect of studying 'up close' a potentially important galaxy formation process.

The tidal forces created by interacting galaxies can literally tear strips from them, producing strikingly long tails of matter⁵ that may become detached and escape into intergalactic space (Fig. 1, overleaf). Curiously, such tails often contain stellar clumps resembling small (dwarf) galaxies — by far the most common type of galaxy. In 1956, Zwicky first suggested that some of these clumps might actually form new galaxies⁶,

and numerical simulations^{7,8} have recently bolstered this idea.

Many dwarf galaxies have irregular shapes, contain lots of interstellar atomic hydrogen (H I) — the raw fuel for making stars — and exhibit unmistakable signs of recent star formation in the form of ionized hydrogen⁹. (This is hydrogen from which the electron has been removed, leaving behind the naked proton.) Clouds of ionized hydrogen are produced by ultraviolet radiation from hot, young stars (10⁶–10⁷ years old), the Orion Nebula in our Galaxy being a local example.

The brightest clumps found in tidal debris — called tidal dwarf galaxies (TDGs) by Braine *et al.* — often share the same properties as irregular dwarf galaxies¹⁰. Circumstantial evidence for a connection seems compelling, yet clouds of molecular hydrogen out of which stars actually form (at least in our Galaxy and others nearby) have not been found in TDGs (Fig. 1). This is embarrassing because gravitationally condensing

H I is expected to become largely molecular (H_2) before forming stars. Star formation is also an inefficient process, so lots of molecular gas should be left over.

By far the most abundant chemical in a molecular cloud is H_2 , but it does not radiate detectable energy at typical cloud temperatures of 20–50 K. Like other astronomers, Braine *et al.* have to infer the amount of H_2 by observing radiation from accompanying trace molecules, principally carbon monoxide (CO). Crucial to this procedure is the assumed ratio of H_2 to CO in the cloud. This is known to depend on the chemical abundance in the surroundings, in this case the TDG. The few existing studies of TDGs suggest that their abundances are sufficiently like those in our Galaxy to justify Braine *et al.*'s use of the 'standard' CO-to- H_2 conversion factor derived for molecular clouds near the Sun.

Many tidal tails have been searched for CO (refs 11,12), but it has been detected in only three^{13–15}. Two serious impediments prevent us from associating these with true TDGs. First, the detected CO is nearby and moving rather slowly with respect to the probable parent galaxies. It is not obviously escaping. Admittedly the situation is unclear because as usual the velocity is measured in only one direction, and the spatial separation relies on only two coordinates. Second, although the CO appears to accompany the atomic gas, the observations do not distinguish between molecules that were pulled out of the parent galaxy or that were formed within the ejected H I. The latter is favoured by observations suggesting that H I in most large galaxies extends farther from the centre of mass than CO, and so should be easier to remove. Such a distinction is vital if one wants the ejected material to look like an irregular dwarf galaxy after leaving the site of the interaction — a journey requiring more than 10^8 years. In the accepted picture, continuing star formation demands ongoing formation of molecular gas.

Braine *et al.*⁴ present the first reasonably strong case that some TDGs can synthesize new molecular clouds. The CO in their clouds is far removed from the parent galaxy and moving rapidly. It also appears to be concentrated at the same location as the H I, and to share its velocity. Coincidence in location and motion is expected if the molecules formed *in situ* from the atomic gas in the runaway cloud, but difficult to understand otherwise. The velocities and distributions of the H I and CO are sufficiently similar to support the case for *in situ* formation. My chief reservation is that the CO emission is sampled at only a few spots — certainly enough to suggest that it follows the atomic gas, but too few to make a solid case. More complete sampling would help.

It is unlikely that all irregular dwarf galaxies were once TDGs — for one thing most

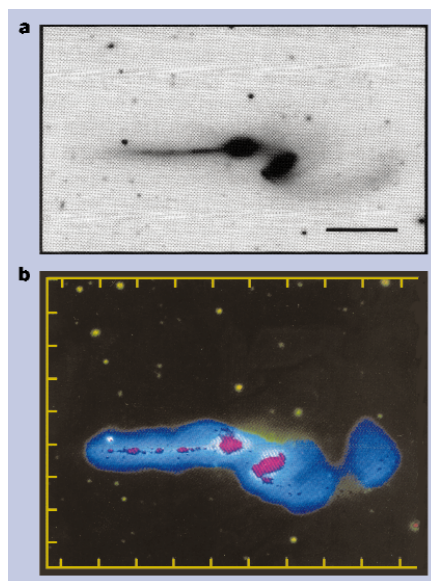


Figure 1 Birthplace of dwarf galaxies?

a, Interacting galaxies in the NGC4676 system, showing the remarkably straight tail of stars pulled from one of the two galaxies¹⁶. The image is a greyscale negative of visible light. The black scale bar shows a distance of 90,000 light years — about the size of our own Galaxy. b, The same system seen with a hydrogen filter, so that atomic hydrogen appears blue, whereas hydrogen that has been ionized (presumably by young stars) appears red. Curiously, the molecular gas expected to accompany these stars is not found in the tail. Images of similar tails analysed by Braine *et al.*⁴ do show signs of molecular hydrogen, suggesting that these tails are involved in continuous star formation.

dwarfs have fewer heavy elements than the TDGs measured so far. To explain even the most element-rich dwarfs, TDGs must survive as long as 10^{10} years — the age of a typical galaxy. The discovery of stars 10^6 – 10^7 years old at the ends of some 10^8 -year-old tidal tails doesn't guarantee long-term stability. Braine *et al.* propose that *in situ* formation of

molecules is a good indicator of stability. I would agree that a gas cloud that is breaking up is not a good place to form molecules. But tidally stripped material is clumpy, probably on smaller scales than we have yet probed. It is possible that molecules could still form within small, stable clumps, however dispersed. Computer simulations of interactions cannot yet resolve this issue; meanwhile I remain sceptical.

One final question: why hasn't CO been seen in other tidal features, such as the merging galaxies shown in Fig. 1? If one assumes the most reasonable value for the CO-to- H_2 conversion factor, then earlier upper limits on detection are generally consistent with the CO signals seen by Braine and colleagues. Again, additional (and more sensitive) data are needed. The detections of molecular gas in TDGs are certain to stimulate such observations, and to lead to a better understanding of the formation of dwarf galaxies. ■

Gary Welch is in the Department of Astronomy and Physics, Saint Mary's University, Halifax, Nova Scotia, B3H 3C3, Canada.

e-mail: gwelch@orion.stmarys.ca

- van den Bergh, S., Abraham, R. G., Ellis, R. S., Tanvir, N. R. & Glazebrook, K. G. *Astron. J.* **112**, 359–368 (1996).
- Balland, C., Silk, J. & Schaeffer, R. *Astrophys. J.* **497**, 541–554 (1998).
- Governato, F., Gardner, J. P., Stadel, J., Quinn, T. & Lake, G. *Astron. J.* **117**, 1651–1656 (1999).
- Braine, J., Lisenfeld, U., Due, P.-A. & Leon, S. *Nature* **403**, 867–869 (2000).
- Toomre, A. & Toomre, J. *Astrophys. J.* **178**, 623–666 (1972).
- Zwicky, F. *Ergeb. Exakt. Naturwiss.* **29**, 344–385 (1956).
- Barnes, J. E. & Hernquist, L. *Nature* **360**, 715–717 (1992).
- Elmegreen, B. G., Kaufman, M. & Thomasson, M. *Astrophys. J.* **412**, 90–98 (1993).
- Hunter, D. A. *Publ. Astron. Soc. Pacif.* **109**, 937–950 (1997).
- Duc, P.-A. & Mirabel, I. F. in *Proc. 186th Symp. Int. Astronomical Union* (eds Barnes, J. E. & Sanders, D. B.) 5 (Kluwer, Dordrecht, 1999).
- Young, J. S., Tacconi, L. J. & Scoville, N. Z. *Astrophys. J.* **269**, 136–146 (1983).
- Smith, B. J. & Higdon, J. L. *Astron. J.* **108**, 837–843 (1994).
- Brouillet, N., Henkel, C. & Baudry, A. *Astron. Astrophys.* **262**, L5–L8 (1992).
- Walter, F. & Heithausen, A. *Astrophys. J.* **519**, L69–L72 (1999).
- Smith, B. J., Struck, C., Kenney, J. D. P. & Jøgee, S. *Astron. J.* **117**, 1237–1248 (1999).
- Hibbard, J. G. & van Gorkum, J. H. *Astron. J.* **111**, 655–695 (1996).

Biodiversity

Extinction by numbers

Stuart L. Pimm and Peter Raven

How large will be the loss of species through human activities? And over what time period might that loss unfold? Habitat destruction is the leading cause of species extinction. Generally, many of the species found across large areas of a given habitat are represented in smaller areas of it. So habitat loss initially causes few extinctions, then many only as the last remnants of habitat are destroyed. Thus, at current rates of habitat destruction, the peak of extinctions might not occur for decades.

But we should not be complacent. On page 853 of this issue, Myers *et al.*¹ document an uneven, highly clumped, distribution of vulnerable species over the world's land surface. Within these 'biodiversity hotspots', habitats are already disproportionately reduced.

Conservatively, there are about seven million species of eukaryote² — a definition encompassing most organisms that would be generally recognized as plants or animals but excluding bacteria, for instance. Most

of these seven million are animals and about 85% are terrestrial.

Humanity is rapidly destroying habitats that are most species-rich. About two-thirds of all species occur in the tropics, largely in the tropical humid forests³. These forests originally covered between 14 million and 18 million square kilometres, depending on the exact definition, and about half of the original area remains⁴. Much of the rain-forest reduction is recent, and clearing now eliminates about 1 million square kilometres every 5 to 10 years⁴⁻⁶. Burning and selective logging severely damages several times the area that is cleared^{5,6}.

To convert habitat loss to species loss, the principles of island ecology are applied to the terrestrial 'islands' that remain in a 'sea' of converted land⁷. The relationship between number of species and island area is nonlinear, and from this one can predict how many species should become extinct as the size of the forest islands shrinks. These doomed species do not disappear immediately, however.

How does one go about calculating the rate of species extinctions from habitat fragments? There have been only a few such estimates, but projections based on a species survivorship curve with a half-life of roughly 50 years seem reasonable⁸. Combining the rate of habitat loss, the species-to-area relationship and the survivorship curve gives a crude extinction curve (curve a in Box 1). From this, we would expect that current extinction rates should be modest — on the order of a thousand species per decade, per

million species, a figure that matches other estimates⁹.

Because the species–area curve is nonlinear, the clearing to date of half of the humid forests is predicted to eliminate only 15% of the species that they contain. The time delays before extinction mean that many more species should be 'threatened' than have already become extinct; that is, they are thought likely to become extinct in the wild in the medium-term future. At least 12% of all plants¹⁰ and 11% of all birds¹¹ come into this category.

Of course, clearing the remaining half of the forests would eliminate the other 85% of species that they contain. The extinction curve should accelerate rapidly to a peak by the middle of the twenty-first century if the rate of forest clearing remains constant. But it will be upon us sooner if that rate is increasing — as seems probable^{4,6}.

Once the extinction peak has passed, the extinction curve declines into the twenty-second century as species are lost from the remaining fragments of habitat. The relative height of the peak depends critically on the amount of habitat that remains. A value of 5% of remaining habitat (see Table 1 on page 854) would protect about 50% of all the forests' species; smaller percentages would lead to smaller estimates of surviving species.

Modest tinkering with parameters does not alter the 'fewer extinctions now, many more later' feature of the extinction curve (curve a in Box 1). But the calculations of Myers *et al.*¹ do. They show that roughly

30–50% of plant, amphibian, reptile, mammal and bird species occur in 25 hotspots that individually occupy no more than 2% of the ice-free land surface (see the map on page 853). That is, terrestrial species with small geographical ranges are numerous and they have highly clumped distributions. Myers *et al.* exclude the oceans from their analysis. But there, too, fishes and other organisms dependent on coral reefs are similarly concentrated¹².

Habitat destruction acts like a cookie cutter stamping out poorly mixed dough⁹. Species found only within the stamped-out area are themselves stamped out. Those found more widely are not. Moreover, species with small ranges are typically scarcer within their ranges than are more widely distributed species, making them yet more vulnerable. Consequently, even random destruction would create centres of extinction that match the concentrations of small-ranged species — the hotspots⁹.

Worse, however, Myers *et al.* show that the cookie cutter is not random — it is malevolent. In the 17 tropical forest areas designated as biodiversity hotspots, only 12% of the original primary vegetation remains, compared with about 50% for tropical forests as a whole. Even within those hotspots, the areas richest in endemic plant species — species that are found there, and only there — have proportionately the least remaining vegetation and the smallest areas currently protected.

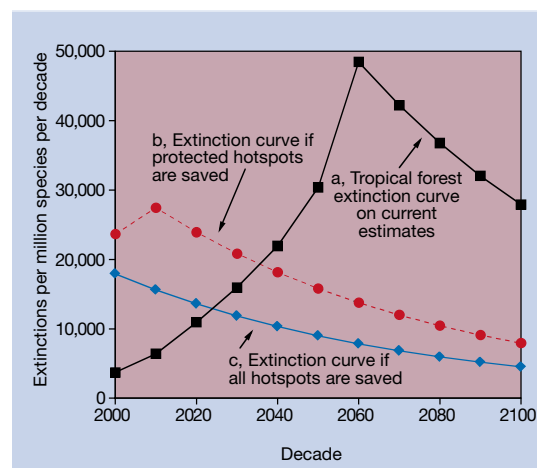
Applying the species–area relationship to the individual hotspots gives the prediction that 18% of all their species will eventually become extinct if all of the remaining habitats within hotspots were quickly protected (curve c in Box 1). Assuming that the higher-than-average rate of habitat loss in these hotspots continues for another decade until only the areas currently protected remain (curve b in Box 1), these hotspots would eventually lose about 40% of all their species. All of these projections ignore other effects on biodiversity, such as the possibly adverse impact of global warming, and the introduction of alien species, which is a well-documented cause of extinction of native species.

Unless there is immediate action to salvage the remaining unprotected hotspot areas, the species losses will more than double. There is, however, a glimmer of light in this gloomy picture. High densities of small-ranged species make many species vulnerable to extinction. But equally this pattern allows both minds and budgets to be concentrated on the prevention of nature's untimely end. According to Myers *et al.*, these areas constitute only a little more than one million square kilometres. Protecting them is necessary, but not sufficient. Unless the large remaining areas of humid tropical forests are also protected, extinctions of those species that are still wide-ranging

Box 1: Extinctions in tropical forests, 2000–2100

Three projections of how numbers of species extinctions in tropical forests may unfold from forest clearance. Curve a is the extinction curve on current estimates, not taking into account biodiversity hotspots. According to the relationship $S_n/S_0 = (A_n/A_0)^{0.25}$ (see refs 6–8), as habitat is reduced from an original area of A_0 to A_n , A_n will hold S_n viable species in year n from an original total of S_0 . The $S_0 - S_n$ doomed species will die off with a half-life of 50 years⁷. With a constant rate of forest clearance, this curve takes time to peak because of the nonlinear relationship between species and area, and the time lags for species to become extinct.

Myers *et al.*¹ identify 25 biodiversity hotspots around the world, of which 17 are in tropical forests. These areas



have already suffered disproportionate loss of primary vegetation, meaning that the many species they contain are under particular threat of extinction. If all remaining habitat in hotspots is saved (as shown in curve c), some 18% of their species

will be lost. The same half-life for currently threatened species is used as in curve a. However, if the hotspots are cleared in the next decade to the point where only currently protected areas are saved (curve b) then the total extinctions will be higher.

S.L.P. & P.R.

should exceed those in the hotspots within a few decades (Box 1).

Stuart L. Pimm is at the Center for Environmental Research and Conservation, MC 5556, Columbia University, 1200 Amsterdam Avenue, New York, New York 10027, USA.

e-mail: stuartpimm@aol.com

Peter Raven is at the Missouri Botanical Garden, PO Box 299, St Louis, Missouri 63166, USA.

e-mail: raven@mobot.org

1. Myers, N., Mittermeier, R. A., Mittermeier, C. G., da Fonseca, G. A. B. & Kent, J. *Nature* **403**, 853–858 (2000).
2. May, R. M. in *Nature and Human Society* (ed. Raven, P. H.) (Natl Acad. Sci. Press, Washington DC, 2000).

3. Raven, P. H. (ed.) *Research Priorities in Tropical Biology* (Natl Acad. Sci. Press, Washington DC, 1980).
4. Skole, D. & Tucker, C. J. *Science* **260**, 1905–1910 (1993).
5. Nepstead, D. C. *et al. Nature* **398**, 505–508 (1999).
6. Cochrane, M. A. *et al. Science* **284**, 1832–1835 (1999).
7. Brooks, T. M., Pimm, S. L. & Collar, N. J. *Conserv. Biol.* **11**, 382–384 (1997).
8. Brooks, T. M., Pimm, S. L. & Oyugi, J. O. *Conserv. Biol.* **13**, 1140–1150 (1999).
9. Pimm, S. L., Russell, G. J., Gittleman, J. L. & Brooks, T. M. *Science* **269**, 347–350 (1995).
10. Walter, K. S. & Gillett, H. J. *IUCN Red List of Threatened Plants* (IUCN, Gland, Switzerland, 1998).
11. Collar, N. J., Crosby, M. J. & Stattersfield, A. J. *Birds to Watch 2* (Smithsonian Inst. Press, Washington DC, 1994).
12. McAllister, D., Schueler, F. W., Roberts, C. M. & Hawkins, J. P. in *Advances in Mapping the Diversity of Nature* (ed. Miller, R.) 155–175 (Chapman & Hall, London, 1994).

Laser physics

Attosecond pulses at last

Paul Corkum

For the past five years, scientists have stood on the threshold of generating attosecond laser pulses but have been unable to cross it. (An attosecond — 10^{-18} s — is to one second as one second is to the age of the Universe.) This may have finally changed with the publication of a paper by Papadogiannis *et al.* (*Phys. Rev. Lett.* **83**, 4289–4292; 1999), who claim to have measured trains of attosecond pulses. The previous record for the shortest laser pulse was 4.5×10^{-15} s (4.5 femtoseconds). Pulses in the femtosecond range led to the development of femtochemistry — making it possible to study chemical reactions in real time — for which the 1999 Nobel Prize in Chemistry was awarded to Ahmed Zewail. But the new science that will ultimately emerge from attosecond research will have its own unique drive.

The approach that Papadogiannis *et al.* use for generating attosecond pulses has been under investigation for some time. They use the short-wavelength harmonics generated when rare gases (such as argon) ionize as a result of irradiation from an intense femtosecond pulse. Harmonics occur at multiples of the frequency of the original femtosecond pulse. Next, the authors select a set of these harmonics, which theory indicates should combine to produce a train of pulses about 100 attoseconds in duration (Fig. 1). Such pulses have probably already been created in many laboratories, but no one has been able to measure them accurately.

Papadogiannis *et al.* may eventually be recognized as the parents of experimental attosecond science because they have actually measured the duration of these pulses. Their measurement process is experimentally simple, but theoretically complex. This is because the production of the attosecond pulses is intrinsically entwined with the measurement. They use a technique influenced by autocorrelation, which is widely

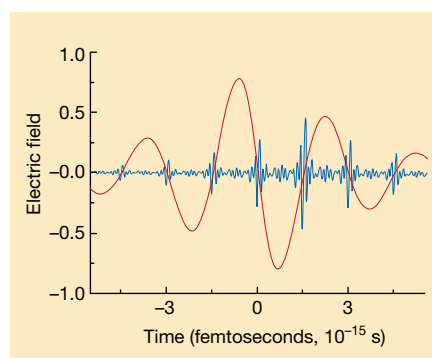


Figure 1 Train of attosecond pulses similar to that produced by Papadogiannis *et al.* Here, the initial femtosecond pulse (red) is much shorter than the one that they used, and the higher-frequency harmonic radiation (blue) is much more intense than in their experiment. The offset between the peak of the initial pulse and of the harmonic radiation illustrates the delay in the harmonic emission imposed by the laser field oscillation.

used for measuring femtosecond pulses and is a close cousin of the pump–probe technique used in femtochemistry.

For autocorrelation measurements, a femtosecond pulse is split into two at a beam splitter. (A beam-splitter functions like a window at night. In a lighted room you can see your reflection in the window while simultaneously being able to see outside.) The two beams are sent through different paths, and usually recombine within a crystal with nonlinear optical properties. Because the harmonic light produced by the nonlinear crystal is stronger when the two pulses are overlapped, observing the signal strength as a function of the difference in the path length of the two beams gives a measurement of the pulse's duration.

Unfortunately, the short duration and wavelength of attosecond pulses means that neither traditional beam splitters nor nonlinear crystals are suitable. Papadogiannis *et al.*

have sidestepped these problems by splitting the femtosecond pulse before the attosecond pulse is produced, and using a rare gas for the dual purpose of producing and measuring the attosecond pulses. All characteristics necessary for measurement are present, but because production and measurement are entwined, their measurement is not completely transparent, so the method is controversial. But the controversy will not last for long because the basic physics behind the measurements is well understood.

Although current research is naturally focused on the production and measurement of attosecond pulses, it is important to look at the future direction of attosecond science. For one thing, it will benefit from the experience gained in previous experiments with ultrashort pulses. This is because we have been performing indirect attosecond experiments (often referred to as strong-field science) for a decade or more and the necessary tools are well developed. For example, normal visible laser pulses contain electric fields that change significantly during 100 attoseconds (Fig. 1). The electric field of the light pulse is proportional to the force that the field exerts on any electrically charged particle. With modern laser technology, the forces can be very large and precisely controlled. So, hidden within the interactions of intense visible laser light with matter are attosecond or near-attosecond phenomena induced by the laser field.

Indeed, indirect attosecond science can explain the attosecond pulses produced by Papadogiannis *et al.* As the electric field of the laser becomes strong, one of the electrons is pulled free from an atom in the argon gas. Once free, it moves in response to the strong force of the laser; first it is driven away from the ion, then back. Its path can be compared to that of a lifeboat launched from a ship in a stormy sea. The ship (or ion) from which it detached is an obstacle that remains in the area and with which it can collide. As with the lifeboat and ship analogy, the wave determines the possible time of collision. In the violent electron–ion collision that may occur, very-short-wavelength radiation can be emitted. So the precise synchronization of the individual attosecond pulses with the much-longer-wavelength radiation that produced them (Fig. 1) is not an accident, but is forced by the field. Quantum mechanics adds ‘fuzziness’ to this essentially classical description, but does not change it much.

No experiments have yet been performed with attosecond pulses, and we cannot even produce an isolated pulse. There are, however, many ideas and proposals waiting in the wings, all involving increasingly precise control over oscillations of the strong laser field. Once attosecond pulses can be produced routinely, indirect and direct attosecond science will become increasingly integrated.

Whereas the goal behind the development

of femtosecond sources was the need to measure the dynamics of molecules and solids, the goal that will drive attosecond science will be control over electronic processes. Electrons are much lighter than atoms and can move much faster, so attosecond pulses will give us a tool for studying the motion of electrons. We can imagine experiments in which atto-

second pulses are used to directly excite electrons, and then the electrons are precisely directed within the molecule by the field oscillations of an intense femtosecond pulse. ■

Paul Corkum is at the Steacie Institute for Molecular Science, National Research Council, 100 Sussex Drive, Ottawa, Ontario, K1A 0R6, Canada. e-mail: paul.corkum@nrc.ca

Neurobiology

Neurofibromin progress on the fly

Ronald L. Davis

One in every 3,500 individuals suffers from the genetic disease neurofibromatosis type 1 (NF1). The many symptoms of the disease, including tumours of the nervous system^{1,2}, result from a deficiency in the protein neurofibromin. Neurofibromin belongs to the family of GTPase-activating proteins (GAPs), which turn off the growth-promoting function of the Ras family of proteins by stimulating the hydrolysis of GTP bound to Ras. Without enough neurofibromin, Ras remains unchecked, resulting in cellular overgrowth and tumours. About half of NF1 patients also suffer from learning disabilities, which have been assumed to result from abnormal brain development due to aberrant neurofibromin control of Ras and cell division. On page 895 of this issue, Guo *et al.*³ offer an alternative explanation. They

show that neurofibromin deficiency in the fruitfly *Drosophila* also causes a learning disability, but that it arises from a requirement for neurofibromin in activating adenylyl cyclase in the mature brain.

The cognitive impairments associated with NF1 are variable. Around 4–8% of NF1 patients are mentally retarded, but the average IQ for NF1 patients (92) is only slightly lower than the population norm. Yet about half of the affected patients have learning disabilities severe enough to require educational assistance^{1,2}. Learning disabilities occur in language-based skills, such as reading, spelling and vocabulary development, and in visuospatial skills, such as face recognition, spatial memory and the judgement of line orientation^{1,2}. In addition, NF1 patients often have problems with the execution of

goal-directed behaviours, including planning, attention and organization (executive function)².

Support for the view^{1,2} that the learning disability arises from a problem in brain development comes from brain scans using magnetic resonance imaging, which in many NF1 patients show between one and five unusual bright spots^{1,2,4}. Although the presence of spots correlates poorly with learning disabilities^{2,5}, they do indicate a physical difference in the brains of some NF1 patients. Reported changes in the number and size of cells called astrocytes in the brain, and the size of certain compartments in the NF1 brain⁴, could also disrupt learning.

Alternatively, neurofibromin might mediate learning through the physiology of the mature brain. The work of Guo *et al.*³ supports this idea. *Drosophila* neurofibromin mutants were tested for their ability to learn an odour paired with an electric shock. The mutants showed impaired learning, like their human counterparts, perhaps because neurofibromin has a similar role in the development or function of insect and mammalian brains.

To test whether the learning deficiency was due to lack of neurofibromin during development or during adulthood, Guo *et al.* created transgenic flies carrying a neurofibromin gene linked to a heat-inducible promoter^{6,7}. This allowed them to turn the gene on in adult mutants, after development was complete. When they did this, the mutants could learn normally. So *Drosophila*

Volcanology

The bulge of Casita

Volcanoes can be dangerous even when dormant. The centre of attention in this photograph lies not in the smoking cone of San Cristobal volcano in the background but in the less-obvious bulge on the nearer slope (circled). The bulge lies on the flanks of Casita, a dormant 1,400-m-high volcano in Nicaragua, some 100 km northwest of the capital, Managua.

Writing in *Geology* (28, 167–170; 2000), Benjamin van Wyk de Vries and colleagues describe their survey of Casita and experimental simulations of how such bulges grow and collapse. The cause of the deformation is hydrothermal activity which, over the course of decades or centuries, can weaken the core edifice of a dormant volcano by turning solid rock into much weaker clays. The resulting structures on the flanks

are unstable, and earthquake activity or heavy rainfall can trigger an avalanche. Indeed, a secondary consequence of bulge formation on Casita was a comparatively small-scale landslide in 1998, which was triggered by Hurricane Mitch and caused widespread devastation.

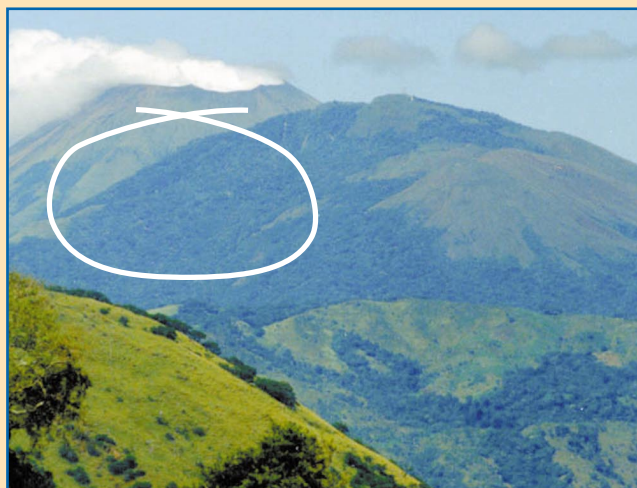
The experiments of van Wyk de Vries *et al.* involved building 10-cm-high cones of sand and plaster to simulate solid rock, with silicone to simulate hydrothermally altered, deformable clays. If the silicone was offset from the central axis, the result was 'slump-like' structures on one side, and eventual collapse of those structures. The process of bulge formation took about 15 minutes. The authors estimate that the bulge on Casita has taken at least 500 years to develop; taking scaling into account, the

experimental result is in broad agreement with that figure.

van Wyk de Vries *et al.* point out that monitoring of dormant volcanoes does not have high priority. But identifying and

keeping tabs on unstable structures such as that on Casita would pay dividends in hazard assessment, especially when large populations are potentially under threat.

Tim Lincoln



B. VAN WYK DE VRIES

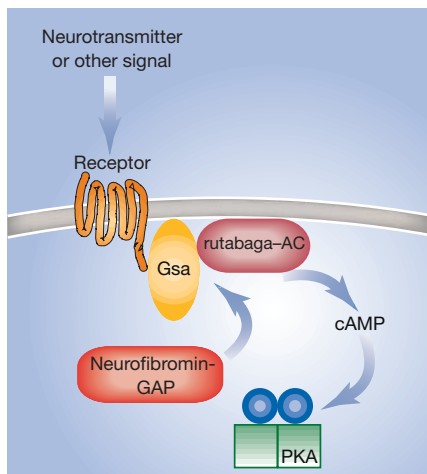


Figure 1 Neurofibromin signalling in learning. G-protein-coupled receptors are linked to adenylyl cyclase (AC) through the α -subunit of the stimulatory G protein, Gs. Receptor activation stimulates adenylyl cyclase to synthesize cyclic AMP (cAMP) and activates the downstream effector, protein kinase A (PKA). Guo *et al.*³ show that the neurofibromin protein is required in this pathway for normal odour learning in *Drosophila*, perhaps by direct interaction with the G protein or with the adenylyl cyclase encoded by the *rutabaga* gene.

neurofibromin is needed in the mature brain for normal odour learning to occur.

But does neurofibromin involvement in learning depend on Ras? In *Drosophila*, the answer seems to be no. It was previously shown that a neuropeptide known as PACAP increases the activity of voltage-gated potassium channels⁸. This response involves the combined activation of the Ras pathway and the cyclic AMP pathway through adenylyl cyclase. Neurofibromin is required for this biological response to the neuropeptide but, surprisingly, the requirement lies not in the Ras pathway but in the cyclic AMP pathway, as application of cyclic AMP itself can restore the response in NF1 mutants⁹. So, in flies, neurofibromin is involved in the activation of adenylyl cyclase (Fig. 1).

Guo *et al.*³ have extended the experimental support for this role in three ways. First, they show that expression in transgenic flies of an active form of protein kinase A, a component of the cyclic AMP pathway, corrects the learning deficiency of NF1 mutants. Second, they find that neurofibromin is required for the normal activation of adenylyl cyclase activity by GTP. Finally, the magnitude of the deficiency in stimulation of adenylyl cyclase by GTP in NF1 mutants is similar to that in learning mutants of the *rutabaga* gene. As *rutabaga* is the structural gene for one type of adenylyl cyclase, it appears that *Drosophila* neurofibromin works specifically through this cyclase.

This adds up to suggest that neurofibromin in the fly has a biochemical function not

yet discovered in mammals: positive regulation of the activation of adenylyl cyclase by GTP. Furthermore, this regulation takes place in mature neurons, and its failure in adults causes learning disabilities. But how might these findings apply to mammalian neurofibromin, NF1 patients and mammalian cognition?

It is now imperative to test whether mammalian neurofibromin can interact with and activate adenylyl cyclase/G-protein complexes. There is plenty of room on the protein for direct interactions, as the GAP domain occupies only 10% of the sequence¹⁰. Furthermore, some alternatively spliced variants of mammalian neurofibromin are expressed only in mature neurons¹¹, offering the potential for isoform-specific functions such as cyclase activation. Nevertheless, even though *Drosophila* neurofibromin is highly homologous to mammalian neurofibromin and has Ras-GAP activity *in vitro*¹², it may be that the fly neurofibromin simply functions differently. Genetic attempts to show that *Drosophila* neurofibromin also negatively regulates Ras have failed¹², perhaps owing to compensation by other GAPs, or perhaps because neurofibromin has nothing to do with Ras signalling in flies.

But can we really learn about human learning disabilities from the fly? After all, the fly has a small brain, limited intellectual capacity, and no higher cognitive traits such as executive function. It seems so, as the conservation between mammalian and *Drosophila* genes observed to date, even for those involved in behaviour, is remarkable. Several genes that are required for *Drosophila* learning, including *dunce*, *rutabaga*, *DCO*, *CREB*, *Volado*^{7,13} and *NF1* (ref. 14), are all involved in mammalian learning or other behaviours. Thus, if the parallels remain true and the learning disabilities of NF1 patients stem from signalling rather than developmental defects, then patient treatment with cognitive enhancers could be approached with optimism.

Ronald L. Davis is in the Department of Molecular and Cellular Biology and the Department of Psychiatry and Behavioral Sciences, Baylor College of Medicine, Houston, Texas 77030, USA.
e-mail: rdavis@bcm.tmc.edu

1. North, K. N. *et al. Neurology* **48**, 1121–1127 (1997).
2. Ozonoff, S. *Am. J. Med. Genet.* **89**, 45–52 (1999).
3. Guo, H.-F., Tong, J., Hannan, F., Luo, L. & Zhong, Y. *Nature* **403**, 895–898 (2000).
4. Gutmann, D. *Arch. Neurol.* **56**, 1322–1323 (1999).
5. Moore, B. D., Slovis, J. M., Schomer, D., Jackson, E. F. & Levy, B. M. *Neurology* **46**, 1660–1668 (1996).
6. Dauwalder, B. & Davis, R. L. *J. Neurosci.* **15**, 3490–3499 (1995).
7. Grotewiel, M. S., Beck, C. D., Wu, K. H., Zhu, X. R. & Davis, R. L. *Nature* **391**, 455–460 (1998).
8. Zhong, Y. *Nature* **375**, 558–592 (1995).
9. Guo, H.-F., Inge, T., Hannan, F., Bernards, A. & Zhong, Y. *Science* **276**, 795–798 (1997).
10. McCormick, F. *Curr. Opin. Genet. Dev.* **5**, 51–55 (1995).
11. Gutmann, D. H., Zhang, Y. & Hirbe, A. *Ann. Neurol.* **46**, 777–782 (1999).
12. The, I. *et al. Science* **276**, 791–794 (1997).
13. Davis, R. L. *Physiol. Rev.* **76**, 299–317 (1996).
14. Silva, A. J. *et al. Nature Genet.* **15**, 281–284 (1997).

Daedalus

Grinding waste away

How to get rid of domestic and urban rubbish? Recycling is troublesome, incineration unpopular, and composting impeded by the presence of plastic and metal. Landfill merely entombs the trash archivally for future archaeologists. One answer might be to feed the stuff to pet goats; they eat anything, and would generate useful metabolic heat from its digestion. Unlike a passive compost heap or landfill, an animal gut is ceaselessly churning. Some animals swallow stones or grit to aid their digestion; these must help the process mechanically or chemically. In this connection, Daedalus recalls that halocarbons can be decomposed by ball-milling them with lime. Mechanical impact and abrasion actually break up the molecules and encourage their reaction.

So Daedalus is designing a mechanically agitated composter for all our rubbish. Its steady churning and pounding, shearing the molecules at the point of impact whenever two chunks of rubbish come together, will degrade its contents at an unprecedented rate. Even plastics will be rapidly oxidized to products edible by the resident bacteria. Chemical, mechanical, thermal and biological decomposition will all proceed in parallel. Too violent an agitation might squash the bacteria faster than they could multiply, but slower rates should be entirely effective. Adding domestic sewage to the mix would solve another disposal problem at the same time.

A large-scale rubbish composter would be a long inclined cylinder rotating on its axis. Rubbish added continuously at the elevated end would work its way slowly down; the cans and bottles in it would act as grinders for the organics. Its inlet zone would be kept anaerobic, to generate useful methane. Downstream, injected air would switch it to aerobic decomposition. Toxins and pollutants would be totally mineralized by this steady mechano-chemical battering. Only rust, glass and ceramic grit would emerge at the far end.

But Daedalus wants to abolish rubbish collection as well. So he plans a simpler domestic composting dustbin, to degrade the rubbish and sewage of one household. It will grind away outside the house, but will be plumbed into the heating system to contribute useful warmth to the domestic interior.

David Jones

The Further Inventions of Daedalus (Oxford University Press), 148 past Daedalus columns expanded and illustrated, is now on sale. Special *Nature* offer: m.curtis@nature.com

Obituary

Paul Sigler (1934–2000)

Paul Sigler died of a heart attack on 11 January. Once met, he was a man whom it was impossible to forget. His zest for science and for life was unmatched.

I was a freshly minted PhD working with David Blow at the MRC Laboratory of Molecular Biology, Cambridge, UK, when in 1964 Sigler joined the group as a PhD student. He was physically imposing, loud, opinionated and always in a rush. One of his first acts was to break off the tips of the glass pipettes so that the liquid would run out more quickly and he could complete his experiments faster. In many ways he qualified as the prototypical Ugly American. In fact, he engendered only warmth and affection. He was smart, a born raconteur and, above all, had unbounded enthusiasm.

Sigler was an undergraduate at Princeton University, trained as a physician at Columbia University in New York, and completed his internship and residency at Columbia-Presbyterian Medical Center. For reasons that are unknown, at least to me, he decided to forgo a career in medicine for one in structural biology. In 1961 he walked into the office of David Davies at the US National Institutes of Health (NIH), introduced himself, and said that he wanted to learn protein crystallography. In Davies' laboratory he showed that pepsin-fluoride could be used both to locate the active site of the enzyme γ -chymotrypsin and to introduce a heavy atom (iodine) which might be used to help determine the enzyme's three-dimensional structure.

A Sigler characteristic was his view that it should be possible to reproduce, then improve, any procedure or result described in the literature. His experimental dexterity, however, did not always match his goals. Sigler himself would recall with relish how, at NIH, he decommissioned two spectrophotometers in succession — the first when, in attempting to control the temperature of the sample, he flooded the optical components with water; the second because, in trying to reduce absorption, he introduced helium, which penetrated into and ruined the photocell.

From the early involvement with γ -chymotrypsin, it was a natural transition to join David Blow, who was working on the structure of its sister enzyme α -chymotrypsin. The timing was impeccable. Before Sigler's arrival, the structure of only a single protein (myoglobin) was known at atomic resolution. Shortly thereafter, in 1965,

Enthusiasm and achievement in structural biology

David Phillips' group in London determined the structure of the second (lysozyme). By the time Sigler left Cambridge he had not only played a key part in the structure determination of α -chymotrypsin, but had also witnessed parallel determinations of two further proteins, ribonuclease and carboxypeptidase. Structural biology had arrived.

After the heady success with α -chymotrypsin, Sigler's move in 1968 to an independent position at the University of Chicago was followed by testing times. He was always attracted to those problems of the greatest biological interest and chose to invest the resources of his laboratory in an assault on the structure of methionyl transfer RNA. It had the promise of a spectacular accomplishment — the first determination of the three-dimensional structure of any RNA molecule. In the event, the groups of Alex Rich at the Massachusetts Institute of Technology and Aaron Klug at the Medical Research Council had superior crystals of the related phenylalanine transfer RNA and reached the goal first. Also, Sigler's determination at Chicago of the structure of trp repressor in complex with DNA was controversial because it lacked direct contacts between the protein and the DNA. As it turned out, however, the work proved to be highly valuable in emphasizing the function of water in mediating protein–DNA interactions.

It was at Yale University, to which he moved in 1989, that Sigler was to have his greatest impact. Support provided by the Howard Hughes Medical Institute, and the infrastructure and environment provided by the Yale structural biology group, allowed him to pursue some of the most enticing problems in biology and led to an extraordinary series of accomplishments.

First, he was able to determine the structure of the protein–DNA complex that is the heart of the machinery that regulates gene expression in all higher organisms. His work, together with that of Stephen Burley at the Rockefeller University, showed for the first time how formation of the complex causes the DNA to be severely bent and unwound. This unprecedented

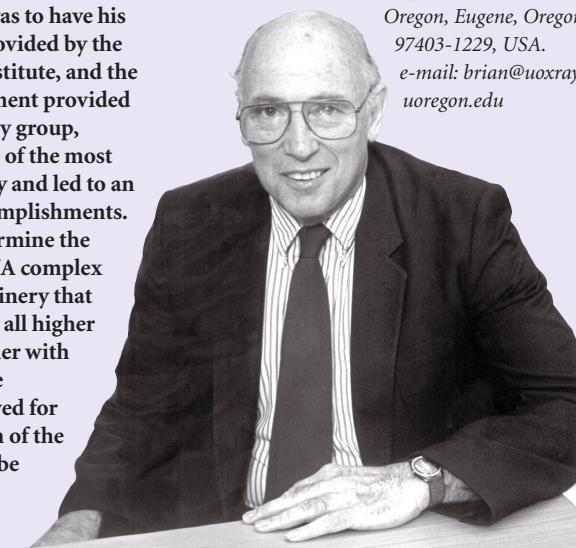
deformation of the DNA underlies the assembly of the overall complex. Related studies also showed how small molecules such as steroid hormones can turn genes on or off.

Second, he showed how a cell can transmit external signals to modulate activities within it. This work centred, in particular, on the way in which light is sensed by the eye. Sigler's determination of the structure of transducin, together with parallel studies of a related protein by Stephen Sprang and Alfred Gilman at the University of Texas Southwestern Medical Center, showed how so-called G proteins mediate signal transduction. This mechanism operates not only in vision but in many biological processes, including smell and the action of stimulants such as adrenaline.

Finally, in a technological *tour de force*, Sigler and his group determined the three-dimensional structure of the bacterial chaperonin GroEL/GroES, a large double-ring structure that forms two back-to-back compartments. Within these chambers, proteins that are being folded to their active shapes can be protected from the crowded environment of the cell. Knowledge of this structure made it possible to see, for the first time, how the process might work (in Sigler's words) "like a two-stroke motor".

At the same time, Sigler could always be relied upon to enliven any meeting or discussion. His large, witty, inquisitive presence invariably left a lasting impression. This is the Paul Sigler we shall remember: demanding yet self-deprecating, competitive yet appreciative of the work of others, a seeker of the big picture yet keen to discuss the results of his most junior colleagues. **Brian W. Matthews**

Brian W. Matthews is in the Howard Hughes Medical Institute and the Institute of Molecular Biology, University of Oregon, Eugene, Oregon 97403-1229, USA.
e-mail: brian@uoxray.uoregon.edu



The sound of many hands clapping

Tumultuous applause can transform itself into waves of synchronized clapping.

An audience expresses appreciation for a good performance by the strength and nature of its applause. The thunder of applause at the start often turns quite suddenly into synchronized clapping, and this synchronization can disappear and reappear several times during the applause. The phenomenon is a delightful expression of social self-organization that provides an example on a human scale of the synchronization processes that occur in numerous natural systems, ranging from flashing Asian fireflies to oscillating chemical reactions^{1–3}. Here we explain the dynamics of this rhythmic applause.

We investigated the mechanism and development of synchronized clapping by making a series of measurements focusing on the collective aspects of the self-organization process as well as on the behaviour of the individuals in the audience. We recorded several theatre and opera performances in Romania and Hungary by using a microphone placed on the ceiling of the hall (Fig. 1a). Typically, after a few seconds of incoherent random clapping, a periodic signal developed (a sign of synchronized clapping), visible in Fig. 1a as pronounced spikes. The transition is also captured by the order parameter (Fig. 1c), which increases as the periodic signal develops, and decreases as it disappears.

Although synchronization increases the strength of the signal at the moment of the clapping, it leads to a decrease in the average noise intensity in the room (Fig. 1d). This is surprising, as the driving force for synchronization would be expected to reflect the desire of the audience to express its enthusiasm by increasing the average noise intensity.

The origin of this conflict between the average noise and synchronization can be understood by correlating the global signal with the behaviour of an individual in the audience. We demonstrated this by recording the local sound intensity in the vicinity of a group of oblivious individuals (Fig. 1b). In the incoherent phase, the local signal was periodic, with a short period corresponding to the fast clapping of an individual in the audience. However, the clapping period suddenly doubled at the start of the synchronized phase (at about 12 seconds in Fig. 1a,b), and slowly decreased as synchronization was lost (Fig. 1e).

The decrease in the average noise intensity is therefore a consequence of the period doubling, because there is less clapping per unit time. An increase in the average noise intensity is possible only by decreasing the

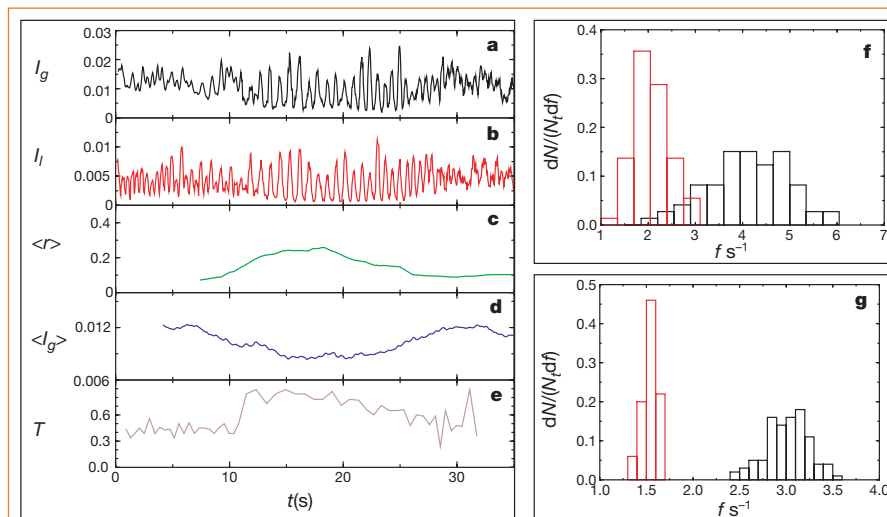


Figure 1 Emergence of synchronization in clapping. **a**, Global noise intensity (I_g) as a function of time. The digitized data were squared and the moving average was determined over a window of size 0.2 s, several times shorter than the clapping period. A characteristic region indicates the appearance and disappearance of the synchronized clapping. Over several performances, we recorded 50 similar sequences of synchronized clapping (for additional data sets and audio recordings, see <http://www.nd.edu/~networks/clap>). **b**, Local noise intensity (I_l), measured by a hidden microphone in the vicinity of a spectator. **c**, Order parameter, r , defined as the maximum of the normalized correlation between the signal $c(t)$ and a harmonic function, $r = [\max_{\phi, T} \int_{-T}^T c(t) \sin(2\pi t/T + \phi) dt] / [\int_{-T}^T c(t) dt]$, where ϕ and T span all possible values. **d**, Average noise intensity, obtained by taking a moving average over a 3-s window of the global noise intensity shown in **a**. **e**, The clapping period, T , defined as the intervals between the clearly distinguishable maxima. **f**, The normalized histogram of clapping frequencies measured for 73 high-school students (isolated from each other) for mode I (black) and mode II (red) clapping. **g**, Normalized histogram for mode I and II clapping obtained for a single student, sampled 100 times over a one-week period.

clapping period, which can indeed occur (Fig. 1e). However, the decreasing clapping period gradually brings the synchronized clapping back to the fast clapping heard in the early asynchronous phase, and synchronization disappears. Apparently, this conflicting desire of the audience simultaneously to increase the average noise intensity and to maintain synchronization leads to the sequence of appearing and disappearing synchronized regimes.

These results indicate that the transition from random to synchronized clapping is accompanied by a period-doubling process. To determine whether period doubling is in fact a necessary condition for synchronization, we investigated the internal frequency of clapping by several individuals in controlled clapping experiments. Individual students, isolated in a room, were instructed to clap as they would immediately after a good performance (mode I clapping) or during the rhythmic applause (mode II clapping).

The frequencies of the two modes of clapping are clearly separated and the average period doubles from mode I to mode II clapping (Fig. 1f). Most important, however, is that the width of the frequency distribution and the relative dispersion of mode II clapping is considerably smaller, a result

that is reproducible for a single individual as well (Fig. 1g).

Our results indicate that after an initial asynchronous phase, characterized by high-frequency clapping (mode I), individuals synchronize by eliminating every second beat, suddenly shifting to a clapping mode with a double period (mode II) where dispersion is smaller. For a group of globally coupled oscillators, the condition for synchronization is that dispersion must be smaller than a critical value^{4,5}. Consequently, period doubling emerges as a condition for synchronization, because it leads to slower clapping modes during which significantly smaller dispersion can be maintained.

Our measurements offer an insight into the mechanism of synchronized clapping: during fast clapping, synchronization is not possible owing to the large dispersion in the clapping frequencies. After period doubling, as mode II clapping with small dispersion appears, synchronization can be and is achieved. However, as the audience gradually decreases the period to enhance the average noise intensity, it slips back to the fast clapping mode with larger dispersion, destroying synchronization.

In summary, the individuals in the audience have to be aware that by doubling their clapping period they can achieve synchro-

nization. This perhaps explains why in the smaller and culturally more homogeneous eastern European communities, synchronized clapping is a daily event, whereas it happens only sporadically in western European and North American audiences. In general, our results offer evidence of a novel route to synchronization, not yet observed in physical or biological systems^{2,3,5,6}.

Z. Nédá*, E. Ravasz*, Y. Brechet†, T. Vicsek‡, A.-L. Barabási§

*Babeş-Bolyai University, Department of Theoretical Physics, Str. Kogălniceanu 1, RO-3400 Cluj-Napoca, Romania

†ENSEEG-LTPCM, INP-Grenoble, Saint-Martin D'Heres, Cedex, France

‡Department of Biological Physics, Eötvös-Loránd University, Budapest 1117, Hungary

§Department of Physics, University of Notre Dame, Notre Dame, Indiana 46556, USA

e-mail: zned@phys.ubbcluj.ro

1. Strogatz, S. H. & Stewart, I. *Sci. Am.* 102–109 (December, 1993).
2. Bottani, S. *Phys. Rev. E* 54, 2334–2350 (1997).
3. Glass, L. & Mackey, M. C. *From Clocks to Chaos: The Rhythms of Life* (Princeton Univ. Press, Princeton, NJ, 1988).
4. Winfree, A. T. *J. Theor. Biol.* 16, 15 (1967).
5. Kuramoto, Y. & Nishikawa, I. *J. Stat. Phys.* 49, 569–605 (1987).
6. Mirolo, R. & Strogatz, S. H. *SIAM J. Appl. Math.* 50, 1645–1662 (1990).

Hormones

Leptin and diabetes in lipoatrophic mice

Lipoatrophic (lipodystrophic) diabetes is a disorder in which insulin resistance and hyperglycaemia are associated with a reduced body-fat mass¹, in contrast to the usual association of diabetes with obesity. Transgenic mice with differing degrees of fat loss can be used as models for lipoatrophy^{2–4}. Using the aP2-SREBP-1c mouse³, which has a moderate fat deficiency, Shimomura *et al.* showed that leptin treatment reverses the diabetes, concluding that insulin resistance in congenital generalized lipodystrophy can be explained by a leptin deficiency⁵. However, we have used a more severe model of lipoatrophy, the A-ZIP/F-1 mouse^{2,6}, in which we find that leptin treatment is only slightly effective in correcting diabetes.

A-ZIP/F-1 mice (Table 1) have an almost complete lack of white adipose (fat) tissue, a severe resistance to insulin, diabetes, and greatly reduced serum leptin levels². We found that infusing leptin into A-ZIP/F-1 mice at the same rate (5 µg day^{−1} for 4 weeks, starting at 7 weeks of age) and to produce the same serum leptin level (3 ng ml^{−1}) as in Shimomura *et al.*'s mice had no effect on serum glucose or insulin concentrations (results not shown). A higher leptin dose (30 µg per day, causing leptin to rise by 5 ng ml^{−1}) did reduce glucose and insulin levels (Fig. 1), food intake (from 6.6 ± 0.3 to 4.7 ± 0.2 g day^{−1}, *P* < 0.001), and liver weight (from 3.04 ± 0.04 to 2.09 ± 0.11 g, *P* < 0.001). Even at this higher dose, however, our mice still had markedly raised blood glucose and insulin levels.

In our A-ZIP/F-1 mice, the efficacy of

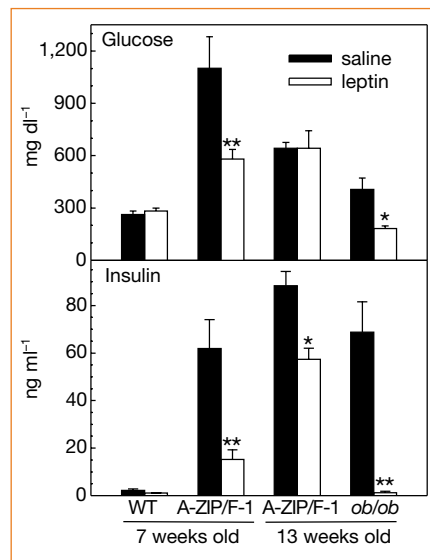


Figure 1 Effect of continuous subcutaneous leptin infusion on serum glucose and insulin concentrations. Leptin (R&D Systems) was infused into male mice by using an osmotic pump (Alzet) at 30 µg day^{−1} for six days, ending at 7 or 13 weeks of age, as indicated. Data are mean and s.e.m. with *n* = 6/group; single and double asterisks indicate differences between treated and control at *P* ≤ 0.01 and *P* < 0.001, respectively. Non-fasting glucose was measured using a Glucometer Elite (Bayer) and insulin was quantified by radioimmunoassay (Linco).

leptin treatment diminished with age: at 13 weeks (the age of Shimomura *et al.*'s mice), leptin had a minimal effect (Fig. 1), and no effect at all at 28 weeks (results not shown). In contrast, leptin infusion into 13-week-old leptin-deficient *ob/ob* mice completely normalized both glucose and insulin levels (Fig. 1).

Evidence from humans and mice supports the conclusion that leptin deficiency cannot completely explain the diabetic phenotype of generalized lipoatrophy. Patients with generalized lipoatrophy are more

prone to diabetes¹ than are those who lack leptin^{7–9}; similarly, A-ZIP/F-1 mice are more diabetic than *ob/ob* mice (Fig. 1). Thus leptin deficiency contributes to the insulin resistance of generalized lipoatrophy, but is neither the sole nor the principal cause of insulin resistance in severe forms of this disease.

The observed differences between the A-ZIP/F-1 and aP2-SREBP-1c mice are probably due to their different amounts of fat, although transgene-specific effects or their different genetic backgrounds may play a part. aP2-SREBP-1c mice have more residual adipose tissue: in these mice, leptin appears to be limiting and its replacement reverses their diabetes. A-ZIP/F-1 mice must experience loss of other functions provided by adipose tissue besides leptin secretion — for example, functions that affect fatty-acid and triglyceride metabolism. Alternatively, adipose tissue might exert a direct or indirect endocrine effect.

Oksana Gavrilova*, Bernice Marcus-Samuels*, Lisa R. Leon*, Charles Vinson†, Marc L. Reitman*

*Diabetes Branch, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland 20892, USA

e-mail: mlr@helix.nih.gov

†Laboratory of Metabolism, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA

1. Foster, D. W. in *Harrison's Principles of Internal Medicine* (eds Isselbacher, K. J. *et al.*) 2131–2136 (McGraw-Hill, New York, 1994).
2. Moitra, J. *et al. Genes Dev.* 12, 3168–3181 (1998).
3. Shimomura, I. *et al. Genes Dev.* 12, 3182–3194 (1998).
4. Burant, C. F. *et al. J. Clin. Invest.* 100, 2900–2908 (1997).
5. Shimomura, I., Hammer, R. E., Ikemoto, S., Brown, M. S. & Goldstein, J. L. *Nature* 401, 73–76 (1999).
6. Gavrilova, O. *et al. Proc. Natl Acad. Sci. USA* 96, 14623–14628 (1999).
7. Montague, C. T. *et al. Nature* 387, 903–908 (1997).
8. Strobel, A., Issad, T., Camoin, L., Ozata, M. & Strosberg, A. D. *Nature Genet.* 18, 213–215 (1998).
9. Farooqi, I. S. *et al. N. Engl. J. Med.* 341, 879–884 (1999).

Shimomura *et al. reply* — Human lipodystrophy (also called lipoatrophic diabetes) is genetically heterogeneous, with the severity of insulin resistance and diabetes mellitus varying widely depending on the degree of reduction in adipose tissue mass and the age of the patient^{1,2}. It is therefore not surprising that two mouse models of lipodystrophy (created by using two different transgenes, A-ZIP/F-1 and aP2-SREBP-1c) vary in their disease severity and in their sensitivity to leptin. The aP2-SREBP-1c animals respond to leptin with a decrease in their insulin and blood sugar levels³, whereas the A-ZIP/F-1 animals of Gavrilova *et al.* apparently manifest leptin resistance. The differences between these two models should not preclude a clinical trial of leptin in leptin-deficient patients with lipodystrophy, with continuation of therapy in those who are leptin-sensitive.

Table 1 Differences in the severity of the A-ZIP/F-1 and aP2-SREBP-1c phenotypes

	Epididymal fat mass (% of wild type)	Brown fat mass (% of wild type)	Glucose (mg dl ^{−1} , non-fasting)	Insulin (ng ml ^{−1} , non-fasting)
A-ZIP/F-1	≤1	~50	~1,000	~60
aP2-SREBP-1c, line A (ref. 3)	~30	~400	~300	~20

Data are for 7-week-old mice.

Ichiro Shimomura*, Robert E. Hammer†, Shinji Ikemoto*, Michael S. Brown*, Joseph L. Goldstein*

*Department of Molecular Genetics, †Department of Biochemistry and Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, Dallas, Texas 75235-9046, USA

1. Foster, D. W. in *Harrison's Principles of Internal Medicine* (eds Fauci, A. S. et al.) 2209–2214 (McGraw-Hill, New York, 1998).
2. Seip, M. & Trygstad, O. *Acta Paediatr.* **413** (suppl.), 2–28 (1996).
3. Shimomura, I., Hammer, R. E., Ikemoto, S., Brown, M. S. & Goldstein, J. L. *Nature* **401**, 73–76 (1999).

Pheromones

Exploitation of gut bacteria in the locust

The congregation of locusts into vast swarms can cause crop devastation of biblical proportions¹. Here we show that guaiacol, a key component of a pheromone derived from locust faecal pellets that promotes the aggregation of locusts^{2–5}, is produced by bacteria in the locust gut. This adaptation by an insect to exploit a common metabolite produced by indigenous gut bacteria has wide implications for our appreciation of the role of the gut microbiota in insects.

Guaiacol (2-methoxyphenol) and phenol are volatile compounds that are both released from the faecal pellets of conventionally reared larval and mature adult desert locusts, *Schistocerca gregaria* (Table 1). Guaiacol production from locusts has previously been attributed to the insect itself. However, desert locusts have a large

but relatively simple gut bacterial biota⁶ which comprises bacterial species acquired from their environment and located in particular in the lining of the hindgut, where faecal pellets form.

We investigated the possible involvement of the gut biota in the production of guaiacol by rearing locusts from surface-sterilized eggs in a sterile isolator system and establishing a breeding colony of axenic (germ-free) locusts by feeding them γ -irradiated freeze-dried grass and bran⁷. Faecal pellets from axenic locusts smelled markedly different from those from locusts with a normal gut biota. Chemical analysis revealed that the difference in odour was due to the absence of guaiacol and low levels of phenol in volatiles released from the germ-free faecal pellets.

We detected guaiacol and phenol in volatiles from the faecal pellets of mono-associated fifth-instar larvae, immature and mature adults carrying a single bacterial species *Pantoea* (= *Enterobacter*) *agglomerans*, a prominent member of the locust-gut biota⁶, and reared on the γ -irradiated diet (Table 1). The smaller amount of volatile phenolics released from young-adult mono-associated insects in comparison with other stages correlates with the lower numbers of bacteria in the gut of newly moulted locusts⁶. These results show that guaiacol originates from the gut bacteria. This finding is supported by experiments demonstrating that three different species of bacteria (including *P. agglomerans*) from the locust gut can produce guaiacol directly from axenic faecal pellets *in vitro*.

The precursor for guaiacol synthesis in faecal pellets must either be a component of plant material or an excretory product of the insect itself. The former is most likely, as the amount of guaiacol produced depends on the diet: more guaiacol was produced by normal locusts fed on fresh wheat seedlings than by those fed freeze-dried γ -irradiated grass. Incubation of locust food with bacteria generated only trace amounts of guaiacol or phenol (data not shown), indicating that digestion of the plant material in the locust gut is required for production of guaiacol by the bacteria.

The most likely precursor for guaiacol synthesis is lignin-derived vanillic acid (4-hydroxy-3-methoxybenzoic acid), which is found in the faeces of both axenic and normal locusts⁸. Microbial transformation of vanillic acid to guaiacol requires a decarboxylation step⁹. Consistent with this, we found guaiacol released by all three species of bacteria from glucose/peptone broth cultures containing vanillic acid (data not shown). Furthermore, faeces from conventionally reared insects fed filter paper impregnated with vanillic acid solution yielded large amounts of guaiacol (Table 1).

The gut bacteria of locusts, as in many

other insects, are considered to be either commensal or facultatively pathogenic¹⁰, and therefore to have little effect on their hosts. Our results show that locusts have adapted to use a pheromonal component that is derived from its digestive waste products by the action of bacteria acquired serendipitously with its food. The gut bacteria also help the locust to defend itself against microbial pathogens, mainly by producing antimicrobial phenolic compounds^{8,11–13}. These contributions by the insect's gut microbiota to its behaviour and survival were previously unsuspected.

Rod J. Dillon, Chris T. Vennard,

A. Keith Charnley

Microbial Pathogenicity Group, Department of Biology and Biochemistry, University of Bath, Bath BA2 7AY, UK
e-mail: r.j.dillon@bath.ac.uk

1. Steedman, A. *Locust Handbook* 2nd edn 2–9 (Overseas Development Natural Resources Institute, London, 1988).
2. Fuzeau-Braesch, S., Genin, E., Jullien, R., Knowles, E. & Papin, C. *J. Chem. Ecol.* **14**, 1023–1033 (1988).
3. Obeng-Ofori, D., Torto, B., Njagi, P. G. N., Hassanali, A. & Amiani, H. *J. Chem. Ecol.* **20**, 2077–2087 (1994).
4. Byers, J. A. *Biol. Rev.* **66**, 347–378 (1991).
5. Pener, M. P. & Yerushalmi, Y. *J. Insect Physiol.* **44**, 365–377 (1998).
6. Hunt, J. & Charnley, A. K. *J. Invertebr. Pathol.* **38**, 378–385 (1981).
7. Charnley, A. K., Hunt, J. & Dillon, R. J. *J. Insect Physiol.* **31**, 477–485 (1985).
8. Dillon, R. J. & Charnley, A. K. *Can. J. Microbiol.* **34**, 1075–1082 (1988).
9. Crawford, R. L. & Olson, P. P. *Appl. Environ. Microbiol.* **36**, 539–543 (1978).
10. Lysenko, O. *Ann. Rev. Microbiol.* **39**, 673–695 (1985).
11. Dillon, R. J. & Charnley, A. K. *J. Invertebr. Pathol.* **47**, 350–360 (1986).
12. Dillon, R. J. & Charnley, A. K. in *The Fungal Spore and Disease Initiation in Plants and Animals* (eds Cole, G. T. & Hoch, H. C.) 129–156 (Plenum, New York, 1991).
13. Dillon, R. J. & Charnley, A. K. *J. Invertebr. Pathol.* **66**, 72–75 (1995).

Pollution

Recovery of breeding success in wild birds

We have found that the breeding success of two insectivorous forest passerines, the great tit *Parus major* and the pied flycatcher *Ficedula hypoleuca*, has markedly improved in the vicinity of a copper-smelting plant during the seven years since it reduced its emissions of heavy metals. Our results demonstrate that reduced pollution loads can positively affect breeding performance of wild bird populations over a relatively short period, even in an area that has suffered decades of heavy-metal pollution.

We collected the data around a copper smelter in Harjavalta (61° 20' N, 22° 10' E) in southwest Finland during 1991–97. Concentrations of heavy metals have increased in the surrounding area of the factory because of long-term deposition^{1–4}. An earlier study of the same area indicated that

Table 1 Volatile phenolic compounds released from locust faecal pellets

Treatment	Guaiacol ($\mu\text{g g}^{-1} \text{d}^{-1}$)	Phenol ($\mu\text{g g}^{-1} \text{d}^{-1}$)
Axenic fifth instar	Not detected	0.5
Axenic mature adult	Not detected	0.3
Monoassociated fifth instar*	6.5	4.3
Monoassociated young adult	1.0	2.0
Monoassociated adult	4.9	8.7
Normal fifth instar† (wheat seedling diet)	44.5	10.7
Normal fifth instar (γ-irradiated grass diet)	4.9	13.1
Normal mature adult (wheat seedling diet)	10.6	12.3
Normal mature adult (filter paper+vanillic acid diet)	38.5	1.4
Normal mature adult (filter paper diet only)	2.6	0.7

Axenic locusts were reared according to ref. 7.

*Monoassociated insects contain the gut bacterium *P. agglomerans* and were fed a γ-irradiated diet.

†Conventional insects contain a normal gut bacterial biota. Volatiles released from locust faeces (derived from > 10 insects per experiment) were analysed by gas chromatography (GC) and the identity of compounds was confirmed by GC-mass spectrometry. The amount of compound was estimated per gram of dry weight of faecal pellets. Further methodological details are available from the authors.

Figure 1 The median lead concentrations ($\mu\text{g g}^{-1}$, dry weight) in femurs of nestlings along the pollution gradient. **a**, *Ficedula hypoleuca*; **b**, *Parus major*. The concentration of lead was measured by atomic absorption spectrophotometry of the femur bones of 6–16-day-old nestlings found dead in their nests or dissected by us. Error bars indicate the third quartile (75%) of the data. Numbers refer to the number of individuals studied. In an ANOVA model with year and distance as independent variables, both year and zone

explained significantly the femur lead concentration (log-transformed data) in both species (*F. hypoleuca*: d.f. (year)=3, $F=17.5$, $P=0.0001$; d.f. (zone)=2, $F=26.7$, $P=0.0001$; *P. major*: d.f. (year)=1, $F=76.8$, $P=0.0001$; d.f. (zone)=2, $F=45.7$, $P=0.0001$).

P. major and *F. hypoleuca* produced fewer fledglings there³, clutch sizes of *F. hypoleuca* were small, and their eggshells were thin and porous when laid in the vicinity of the pollution source⁴. These poor-quality eggshells were linked to large amounts of heavy metals consumed by the birds in their invertebrate prey and with the paucity of calcium-rich food in the acidified forests^{3,5}.

Emission of sulphur dioxide decreased by about 66%, particulate dust by 63%, and lead by 95% during 1990–97. We analysed

seven years' breeding results to determine whether there had been any change in breeding parameters during this period. Our analysis covered two periods, 1991–93 and 1994–97, five zones (<1, 1–2, 2–4, 4–8 and >8 km from the factory) and five habitat types, from the most barren to the most luxuriant, according to the ground-layer vegetation.

We found that lead concentrations in nestlings decreased in the vicinity of the factory from 1991 to 1996 by about 90% in

F. hypoleuca and 87% in *P. major* (Fig. 1). The clutch size of *F. hypoleuca* increased significantly at points closest (<1 km) to the reduced-pollution source, but remained the same at more distant sites (Fig. 2a). The clutch size of *P. major* failed to show a pollution-related decrease along the pollution gradient in either period, and there were no significant changes between the two periods (Fig. 2c).

The number of fledglings increased by about 1.6 chicks per nest at points close to the source in both bird species during 1991–93 and 1994–97 (Fig. 2b,d), reflecting partly the increased clutch size of *F. hypoleuca*. We found no significant effects of habitat on clutch size or on fledgling number in either species, so any changes in habitat type along the gradient do not bias the results. The relative fledgling production (fledglings in the furthest zone/fledglings in the nearest zone) correlated negatively and significantly with yearly emissions of lead ($\log 1,000 \text{ kg yr}^{-1}$) in *F. hypoleuca* ($r_s = -0.83$, $P=0.042$, $n=6$ years; we omitted 1994 because of insufficient data) and was marginally significant in *P. major* ($r_s = -0.71$, $P=0.071$, $n=7$ years).

Both bird species probably benefited from the recovery of forest vegetation. The decreased concentration of sulphuric oxides and heavy metals in tree leaves and needles probably promotes the recovery of herbivorous insect populations at the polluted sites^{6,7}. This is an important change for foliage-gleaning bird species that rely strongly on tree-living insect larvae for feeding their nestlings⁸.

Our results indicate that bird populations can recover relatively rapidly in an area that has suffered long-term pollution from heavy metals, as evidenced by the larger clutches of *F. hypoleuca* and the increase in fledgling production for both bird species by more than one chick during the seven-year period. The heavy-metal residues found in birds, as well as their breeding parameters, seem to reflect the level of current emissions⁹ more closely than heavy-metal accumulation in the soil over the past decades.

Tapio Eeva, Esa Lehtikoinen

Section of Ecology, Department of Biology,
University of Turku, FIN-20014 Turku, Finland
e-mail: teevea@utu.fi

- Hynninen, V. *Ann. Bot. Fenn.* **23**, 83–90 (1986).
- Fritze, H., Niini, S., Mikkola, K. & Mäkinen, A. *Biol. Fertil. Soils* **8**, 87–94 (1989).
- Eeva, T. & Lehtikoinen, E. *Oecologia* **108**, 631–639 (1996).
- Eeva, T. & Lehtikoinen, E. *Oecologia* **102**, 312–323 (1995).
- Graveland, J., van der Wal, R., van Balen, J. H. & van Noordwijk, A. J. *Nature* **368**, 446–448 (1994).
- Jussila, I. *SYKESarja B* **12**, 1–78 (1997).
- Heliövaara, K. & Väisänen, R. in *Acidification in Finland* (eds Kauppi, P., Anttila, P. & Kenttämies, K.) 447–467 (Springer, New York, 1990).
- Eeva, T., Lehtikoinen, E. & Pohjalainen, T. *Ecology* **78**, 1120–1131 (1997).
- Nyholm, N. E. I. *Arch. Environ. Contam. Toxicol.* **35**, 632–637 (1998).

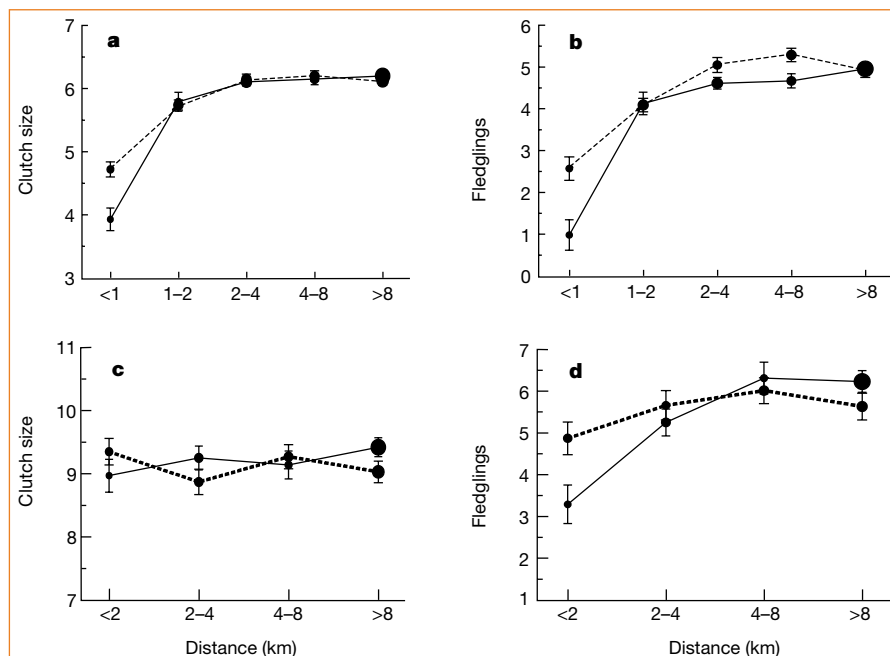


Figure 2 Clutch sizes and fledgling numbers at different distances from the pollution source. Least-square means ($\pm$ s.e.) for the **a**, clutch size and **b**, fledgling number of *Ficedula hypoleuca* (n represents 1,562 clutches and 1,380 broods, respectively); and **c**, clutch size and **d**, fledgling number of *Parus major* (n represents 651 clutches and 573 broods) at five zones around the pollution source in 1991–93 (full line), and in 1994–97 (dashed line). Symbol size denotes the proportional number of nests in each zone. For *P. major*, distance zones I and II were combined in the analysis because of the small number of nests in zone I. In ANOVA models in which the time period, distance and habitat type were independent variables, the temporal change near the pollution source (time $\times$ zone) was significant for the clutch size and fledgling number of *F. hypoleuca* (d.f. = 4, $F=3.9$, $P=0.0041$; d.f. = 4, $F=3.9$, $P=0.0036$, respectively) and for fledgling number of *P. major* (d.f. = 4, $F=4.5$, $P=0.0041$), but non-significant for the clutch size of *P. major* (d.f. = 4, $F=2.2$, $P=0.091$).

Biodiversity hotspots for conservation priorities

Norman Myers*, Russell A. Mittermeier†, Cristina G. Mittermeier†, Gustavo A. B. da Fonseca‡ & Jennifer Kent§

* Green College, Oxford University, Upper Meadow, Old Road, Headington, Oxford OX3 8SZ, UK

† Conservation International, 2501 M Street NW, Washington, DC 20037, USA

‡ Centre for Applied Biodiversity Science, Conservation International, 2501 M Street NW, Washington, DC 20037, USA

§ 35 Dorchester Close, Headington, Oxford OX3 8SS, UK

Conservationists are far from able to assist all species under threat, if only for lack of funding. This places a premium on priorities: how can we support the most species at the least cost? One way is to identify 'biodiversity hotspots' where exceptional concentrations of endemic species are undergoing exceptional loss of habitat. As many as 44% of all species of vascular plants and 35% of all species in four vertebrate groups are confined to 25 hotspots comprising only 1.4% of the land surface of the Earth. This opens the way for a 'silver bullet' strategy on the part of conservation planners, focusing on these hotspots in proportion to their share of the world's species at risk.

The number of species threatened with extinction far outstrips available conservation resources, and the situation looks set to become rapidly worse^{1–4}. This places a premium on identifying priorities. How can we protect the most species per dollar invested? This key question is at the forefront of conservation planning, and forms the focus of this article. By concentrating on areas where there is greatest need and where the payoff from safeguard measures would also be greatest, conservationists can engage in a systematic response to the challenge of large-scale extinctions ahead.

A promising approach is to identify 'hotspots', or areas featuring exceptional concentrations of endemic species and experiencing exceptional loss of habitat^{5–9}. Here we focus on species, rather than populations or other taxa, as the most prominent and readily recognizable form of biodiversity. This is not to suggest that

populations and even ecological processes are not important manifestations of biodiversity, but they do not belong in this assessment. There are other types of hotspot^{10,11}, featuring richness of, for example, rare^{12,13} or taxonomically unusual species^{14,15}. This article considers only hotspots as defined above. Concentrating a large proportion of conservation support on these areas would go far to stem the mass extinction of species that is now underway.

The hotspots' boundaries have been determined by 'biological commonalities'. Each of the areas features a separate biota or community of species that fits together as a biogeographic unit. This is apparent in the case of islands or island groups such as New Caledonia, New Zealand, the Caribbean, Polynesia/Micronesia, Madagascar and the Philippines. Much the same applies to 'ecological islands' in clearly defined continental units such as the Cape

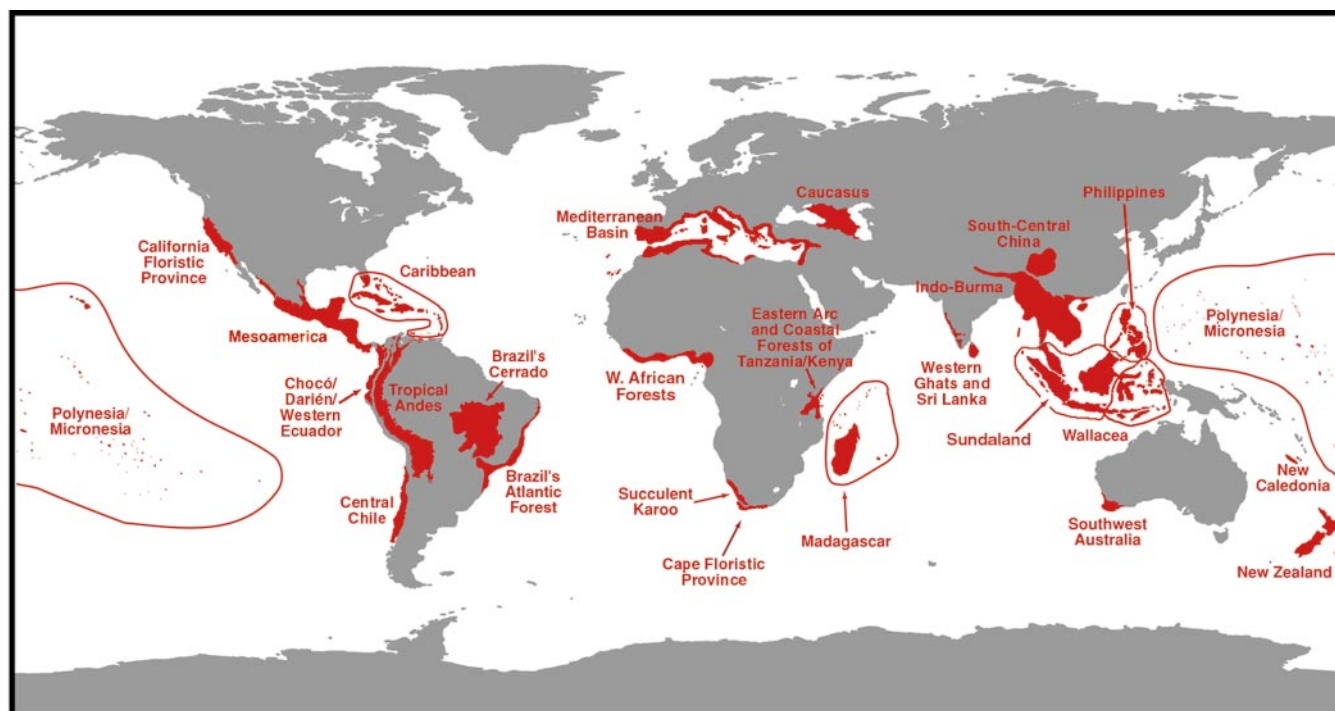


Figure 1 The 25 hotspots. The hotspot expanses comprise 30–3% of the red areas.

Floristic Province, the Eastern Arc and Coastal Forests of Tanzania/Kenya (hereafter abbreviated to 'Eastern Arc'), southwestern Australia and Caucasus. In other areas the definition of a hotspot's boundaries derives from recognized divisions such as Wallace's line between Sundaland and Wallacea, or the Kangar–Pattani line between Indo-Burma and Sundaland. In still other areas, the definition reflects a best-judgement opinion from experts in the field. Were larger hotspots, for example, the Tropical Andes, Mesoamerica, Indo-Burma and Sundaland to be subdivided into areas the size of the smaller hotspots, they would still meet the criterion of biological commonalities; and the result would be a far larger number of mini-hotspots, making for a much more complicated assessment and diffusing the essential strategy of just 25 hotspots designated for priority conservation.

This article is a qualitative as well as a quantitative advance on a preliminary effort^{5,6}, which limited itself to vascular plants in 18 hotspots. The number of hotspots has been increased to 25. More importantly, the expanded criteria require that a hotspot contains endemic plant species comprising at least 0.5% of all plant species world-wide. Here we include four categories of vertebrate species, bringing the number of endemics to almost three times more than in the earlier papers. We analyse key questions of species/area ratios and congruence among taxa. Finally, we present a way to determine the hottest hotspots and thus to pinpoint super priorities.

Analytic methods

The basic analysis is driven by two criteria: species endemism and degree of threat. The main source of data for both plants and vertebrates has been more than 100 scientists with abundant experience in countries concerned and around 800 references in the professional literature (see Supplementary Information). Additional details are available in ref. 16; supplementary sources on plants include refs 17–19. The species dimension is based in the first instance on vascular plants (comprising around 90% of all plants, and hereafter referred to as 'plants'), as they are essential to virtually all forms of animal life and are fairly well known scientifically. To qualify as a hotspot, an area must contain at least 0.5% or 1,500 of the world's 300,000 plant species²⁰ as endemics. In fact,

15 of the 25 hotspots contain at least 2,500 endemic plant species, and 10 of them at least 5,000.

The four vertebrate groups, mammals, birds, reptiles and amphibians, comprise 27,298 species, consisting of 4,809 mammals²¹, 9,881 birds²², 7,828 reptiles²³ and 4,780 amphibians²⁴. The other vertebrate group, fishes, is excluded because data are generally poor (there could well be at least 5,000 species waiting to be discovered²⁵, or more than all mammals). Hereafter 'vertebrates' refers to all vertebrates except fishes. Vertebrates do not serve as an alternative determinant of hotspot status, nor do their endemics have to comprise 0.5% of global totals. If an area qualifies by the 0.5% plants criterion (and the habitat threat criterion), it makes the list. Vertebrates serve as back-up support, and also to determine congruence and to facilitate other comparisons among the hotspots.

The analysis omits invertebrates, which are largely undocumented but probably make up at least 95% of all species, the bulk of them insects. To the extent that the five categories of endemic species assessed are sometimes matched by similar concentrations of endemic insect species, the hotspots thesis can be applied to invertebrates as well. In any case, if we were to lose, say, half of endemic plant species, we could well lose a large and perhaps similar proportion of insect species. The fig genus, for example, being the most widespread of plant genera in the tropics, comprises more than 900 species, each of which is pollinated by a single wasp species; conversely, the wasps depend on the figs' ovaries as sites for their larvae to develop²⁶. Although the plant/insect connection is variable in general application^{27–30}, it is supported by the many pollination, herbivory and other relationships between plants and insects.

The endemism data tend to be minimalist for two reasons. One is the lack of recent documentation in the form of, for example, modern floras. For instance, there is no up-to-date account of Brazil's plant species even though the country is believed to harbour the Earth's richest flora, at least 50,000 species or one-sixth of the planetary total. Second, and more importantly, endemism data almost always relate only to individual countries or parts of countries, whereas 12 of the hotspots extend across two or more countries and six across four or more countries. In these cases, it has been difficult to compute regional totals for hotspot-wide endemics,

Table 1 The 25 hotspots

Hotspot	Original extent of primary vegetation (km ²)	Remaining primary vegetation (km ²) (% of original extent)	Area protected (km ²) (% of hotspot)	Plant species	Endemic plants (% of global plants, 300,000)	Vertebrate species	Endemic vertebrates (% of global vertebrates, 27,298)
Tropical Andes	1,258,000	314,500 (25.0)	79,687 (25.3)	45,000	20,000 (6.7%)	3,389	1,567 (5.7%)
Mesoamerica	1,155,000	231,000 (20.0)	138,437 (59.9)	24,000	5,000 (1.7%)	2,859	1,159 (4.2%)
Caribbean	263,500	29,840 (11.3)	29,840 (100.0)	12,000	7,000 (2.3%)	1,518	779 (2.9%)
Brazil's Atlantic Forest	1,227,600	91,930 (7.5)	33,084 (35.9)	20,000	8,000 (2.7%)	1,361	567 (2.1%)
Choc/Darien/Western Ecuador	260,600	63,000 (24.2)	16,471 (26.1)	9,000	2,250 (0.8%)	1,625	418 (1.5%)
Brazil's Cerrado	1,783,200	356,630 (20.0)	22,000 (6.2)	10,000	4,400 (1.5%)	1,268	117 (0.4%)
Central Chile	300,000	90,000 (30.0)	9,167 (10.2)	3,429	1,605 (0.5%)	335	61 (0.2%)
California Floristic Province	324,000	80,000 (24.7)	31,443 (39.3)	4,426	2,125 (0.7%)	584	71 (0.3%)
Madagascar*	594,150	59,038 (9.9)	11,548 (19.6)	12,000	9,704 (3.2%)	987	771 (2.8%)
Eastern Arc and Coastal Forests of Tanzania/Kenya	30,000	2,000 (6.7)	2,000 (100.0)	4,000	1,500 (0.5%)	1,019	121 (0.4%)
Western African Forests	1,265,000	126,500 (10.0)	20,324 (16.1)	9,000	2,250 (0.8%)	1,320	270 (1.0%)
Cape Floristic Province	74,000	18,000 (24.3)	14,060 (78.1)	8,200	5,682 (1.9%)	562	53 (0.2%)
Succulent Karoo	112,000	30,000 (26.8)	2,352 (7.8)	4,849	1,940 (0.6%)	472	45 (0.2%)
Mediterranean Basin	2,362,000	110,000 (4.7)	42,123 (38.3)	25,000	13,000 (4.3%)	770	235 (0.9%)
Caucasus	500,000	50,000 (10.0)	14,050 (28.1)	6,300	1,600 (0.5%)	632	59 (0.2%)
Sundaland	1,600,000	125,000 (7.8)	90,000 (72.0)	25,000	15,000 (5.0%)	1,800	701 (2.6%)
Wallacea	347,000	52,020 (15.0)	20,415 (39.2)	10,000	1,500 (0.5%)	1,142	529 (1.9%)
Philippines	300,800	9,023 (3.0)	3,910 (43.3)	7,620	5,832 (1.9%)	1,093	518 (1.9%)
Indo-Burma	2,060,000	100,000 (4.9)	100,000 (100.0)	13,500	7,000 (2.3%)	2,185	528 (1.9%)
South-Central China	800,000	64,000 (8.0)	16,562 (25.9)	12,000	3,500 (1.2%)	1,141	178 (0.7%)
Western Ghats/Sri Lanka	182,500	12,450 (6.8)	12,450 (100.0)	4,780	2,180 (0.7%)	1,073	355 (1.3%)
SW Australia	309,850	33,336 (10.8)	33,336 (100.0)	5,469	4,331 (1.4%)	456	100 (0.4%)
New Caledonia	18,600	5,200 (28.0)	526.7 (10.1)	3,332	2,551 (0.9%)	190	84 (0.3%)
New Zealand	270,500	59,400 (22.0)	52,068 (87.7)	2,300	1,865 (0.6%)	217	136 (0.5%)
Polynesia/Micronesia	46,000	10,024 (21.8)	4,913 (49.0)	6,557	3,334 (1.1%)	342	223 (0.8%)
Totals	17,444,300	2,122,891 (12.2)	800,767 (37.7)	†	133,149 (44%)	†	9,645 (35%)

Documentation of plant and vertebrate species and endemism can be found in Supplementary Information.

* Madagascar includes the nearby islands of Mauritius, Reunion, Seychelles and Comores.

† These totals cannot be summed owing to overlapping between hotspots.

and we have often had to depend on best-judgement estimates by over 100 scientists with abundant experience in the countries concerned. In a few instances, we have had to accept a simple summation of country-by-country totals, which surely underestimates regional totals. To this extent, many of the endemism estimates are distinctly conservative.

A second determinant of hotspot status, applied only after an area has met the 'plants' criterion, is the degree of threat through habitat loss. To qualify, a hotspot should have lost 70% or more of its primary vegetation, this being the form of habitat that usually contains the most species, especially endemics. Eleven hotspots have already lost at least 90% and three have lost 95%. The 70% cutoff is justified on the grounds that most large-scale concentrations of endemic plant species occur within the 25 hotspots as delineated. Other concentrations of plant endemics with perhaps another 15% of the Earth's plant species occur in three regions designated as 'major tropical forest wilderness areas', each retaining 75% of its primary vegetation (see below). There are few other areas with comparable concentrations. Moreover, were the 70% cutoff to be replaced with 60%, this would admit hardly any other hotspots, whereas a 90% cutoff would exclude 11 of the hotspots.

Finally, the analysis is limited to the terrestrial realm (Conservation International is preparing an analysis of marine species and conservation priorities).

The area-by-area findings are presented in Tables 1–6 and Fig. 1. For further information regarding the sources of our statistics, see the list of references and experts in Supplementary Information.

There is variability in the precision and accuracy of data. This is to be expected given the range of areas and the degree of documentation available. In many instances, the statistical information is considered to be accurate to within 5%. In most others, it is sufficiently accurate to rank as sound support for working estimates. For example, the Tropical Andes is believed to contain at least 20,000 known plant endemics, this being a rounded figure (many more species, probably thousands, remain to be discovered there). Another 14 such totals are rounded. The Cape Floristic Province, by contrast, is considered to contain exactly 5,682 known plant endemics; the same precision applies to another nine hotspots. Similar considerations apply to vertebrate data and to estimates of remaining primary vegetation.

This overall approach, uneven as it is, is justified for an analysis that seeks to convert a profound problem into a fine opportunity. After all, to decide that a potential hotspot should not be evaluated because it lacks a conventional degree of accurate data is effectively to decide that its conservation needs cannot be evaluated either, in which case its cause tends to go by default. Uncertainty can cut both ways.

Chief findings

The 25 hotspots contain the remaining habitats of 133,149 plant species (44% of all plant species world-wide; Table 1) and 9,645 vertebrate species (35%; Table 2). These endemics are confined to an aggregate expanse of 2.1 million square kilometres, or 1.4% of the Earth's land surface. They formerly occupied 17.4 million square kilometres or 11.8% of the Earth's land surface. They are so threatened that, having already lost an aggregate of 88% of their primary vegetation, they all seem likely, in the absence of greatly increased conservation efforts, to lose much if not most of their remaining primary vegetation within the foreseeable future.

The 25 hotspots feature several habitat types at global scale. Predominant are tropical forests, appearing in 15 hotspots, and Mediterranean-type zones, in five. Nine are mainly or completely made up of islands; almost all tropical islands fall into one or another hotspot. Sixteen hotspots are in the tropics, which largely means developing countries where threats are greatest and conservation resources are scarcest.

Leading hotspots

Some hotspots are much richer than others in terms of their numbers of endemics (Table 3). (Three other modes of comparison are presented below.) Each of five hotspots—the Tropical Andes, Sundaland, Madagascar, Brazil's Atlantic Forest and the Caribbean—contains endemic plants and vertebrates amounting to at least 2% of total species world-wide. Together, they comprise 20% and 16%, respectively, of all plants and vertebrates, and 45% of all the hotspots' endemic plants and vertebrates alike, but they comprise a mere 0.4% of the Earth's land surface. At the same time, they feature some of the most depleted habitats: the Caribbean retains only 11.3% of its primary vegetation, Madagascar 9.9%, Sundaland 7.8% and Brazil's Atlantic Forest 7.5%. These five hotspots, with

Table 2 Vertebrate species and endemism

Hotspot	Bird species and endemism		Mammal species and endemism		Reptile species and endemism		Amphibian species and endemism		Total species and endemism	
Tropical Andes	1,666	677	414	68	479	218	830	604	3,389	1,567
Mesoamerica	1,193	251	521	210	685	391	460	307	2,859	1,159
Caribbean	668	148	164	49	497	418	189	164	1,518	779
Brazil's Atlantic Forest	620	181	261	73	200	60	280	253	1,361	567
Choco/Darien/W. Ecuador	830	85	235	60	210	63	350	210	1,625	418
Brazil's Cerrado	837	29	161	19	120	24	150	45	1,268	117
Central Chile	198	4	56	9	55	34	26	14	335	61
California Floristic Province	341	8	145	30	61	16	37	17	584	71
Madagascar	359	199	112	84	327	301	189	187	987	771
Eastern Arc and Coastal Forests of Tanzania/Kenya	585	22	183	16	188	50	63	33	1,019	121
West African Forests	514	90	551	45	139	46	116	89	1,320	270
Cape Floristic Province	288	6	127	9	109	19	38	19	562	53
Succulent Karoo	269	1	78	4	115	36	10	4	472	45
Mediterranean Basin	345	47	184	46	179	110	62	32	770	235
Caucasus	389	3	152	32	76	21	15	3	632	59
Sundaland	815	139	328	115	431	268	226	179	1,800	701
Wallacea	697	249	201	123	188	122	56	35	1,142	529
Philippines	556	183	201	111	252	159	84	65	1,093	518
Indo-Burma	1,170	140	329	73	484	201	202	114	2,185	528
South Central China	686	36	300	75	70	16	85	51	1,141	178
Western Ghats/Sri Lanka	528	40	140	38	259	161	146	116	1,073	355
SW Australia	181	19	54	7	191	50	30	24	456	100
New Caledonia	116	22	9	6	65	56	0	0	190	84
New Zealand	149	68	3	3	61	61	4	4	217	136
Polynesia/Micronesia	254	174	16	9	69	37	3	3	342	223
Total endemics and % of global total	*	2,821	*	1,314	*	2,938	*	2,572	9,645	
		28.5%		27.3%		37.5%		53.8%	35.3%	

* These totals cannot be summed owing to overlapping between hotspots.

four others, contain endemics amounting to 30.1% and 25.0% of the global totals for plant and vertebrate species, respectively, in 0.7% of the Earth's land surface.

Some hotspots are likewise significant in having their endemic species concentrated in exceptionally small areas (Table 4). The Eastern Arc contains 1,500 endemic plants in 2,000 square kilometres, giving a ratio of 75 species to 100 square kilometres, represented as 75:1, and 121 endemic vertebrates for a ratio of 6.1:1, both ratios topping the lists for all hotspots. Similarly, New Caledonia, with 5,200 square kilometres, works out at 49:1 and 1.6:1, and the Philippines with 9,023 square kilometres at 64.7:1 and 5.7:1. The rest range from 33.3:1 to 1.2:1 for plants and 2.9:1 to 0.03:1 for vertebrates.

Congruence among species categories

In several hotspots there is species congruence insofar as high counts for endemic plants are matched by high counts for endemic vertebrates (Table 5). (For analysis of congruence in other areas, see refs 12 and 31.) This factor reinforces the conservation priority thesis, especially in those hotspots with the most endemic species (Table 3). There can also be high congruence in areas with lower species counts, for example, 80% in the Eastern Arc with 0.5% of plant species and 0.4% of vertebrate species.

Endemic plants in the Tropical Andes comprise 6.7% of all plant species world-wide, and its endemic vertebrates 5.7%, with 85% congruence; Madagascar's species comprise 3.2% and 2.8%, respectively, with 88% congruence; and the Caribbean's 2.3% and 2.9%, with 79%. (The first is a large area where one could expect high congruence; the other two are only one-fifth and one-tenth as big, respectively.) In contrast, Cape Floristic Province possesses 1.9% of all plants but only 0.2% of all vertebrates, for 11% congruence, and the Mediterranean Basin possesses 4.3% of all plants but only 0.9% of all vertebrates, for 21%. Congruence tends to be high in tropical forest hotspots, and generally low in Mediterranean-type hotspots and other drier areas with their meagre counts for endemic vertebrates.

The hottest hotspots

The analysis so far has considered five key factors: numbers of endemics and endemic species/area ratios for both plants and vertebrates, and habitat loss. These factors do not carry equal weight, so they cannot be combined into a single quantitative ranking. For comparative purposes in qualitative fashion, Table 6 lists the eight 'hottest hotspots', which appear at least three times in the top ten listings for each factor. The leaders are Madagascar, the Philippines and Sundaland, appearing for all five factors, followed by Brazil's Atlantic Forest and the Caribbean, appearing for four. Three of these hotspots, Madagascar, the Philippines and the Caribbean, have small areas, which further highlights their importance.

Table 3 Leading hotspots in terms of endemics

Hotspot	Endemic plants (% of global total, 300,000)	Endemic vertebrates (% of global total, 27,298)
Tropical Andes*	20,000 (6.7)	1,567 (5.7)
Sundaland*	15,000 (5.0)	701 (2.6)
Madagascar*	9,704 (3.2)	771 (2.8)
Brazil's Atlantic Forest*	8,000 (2.7)	567 (2.1)
Caribbean*	7,000 (2.3)	779 (2.9)
Sub-totals (% rounded)	59,704 (19.9)	4,385 (16.1)
Mesoamerica	5,000 (1.7)	1,159 (4.2)
Mediterranean Basin	13,000 (4.3)	235 (0.9)
Indo-Burma	7,000 (2.3)	528 (1.9)
Philippines	5,832 (1.9)	519 (1.9)
Totals	90,536 (30.1)†	6,826 (25.0)

* Hotspots with at least 2% of the world's endemic plants and vertebrates, and comprising only 0.4% of the Earth's land surface (all nine amount to 0.7% of the Earth's land surface).

† This would total 30.2% but for rounding of numbers in the individual hotspots.

Two additional hotspots, the Tropical Andes and the Mediterranean Basin, should be considered as hyper-hot candidates for conservation support in light of their exceptional totals of endemic plants: 20,000 and 13,000, respectively. The Tropical Andes is at the top for endemic vertebrates too, and the Mediterranean third after Sundaland for endemic plants, with 34% more than the fourth hotspot. But they do not rank in more than two of the five factor listings. Similarly, Mesoamerica is second for endemic vertebrates (49% more than the third highest), but it scores only tenth for endemic plants.

Higher taxa assessment

The analysis can be complemented by an assessment of endemism among higher taxa such as families and genera. Madagascar (including nearby Indian Ocean islands) possesses 11 endemic families and 310 endemic genera of plants, 5 endemic families and 14 endemic genera of primates, and 5 endemic families and 35 endemic genera of birds. Cape Floristic Province has 6 endemic families and 198 endemic genera of plants; and New Caledonia has 5 endemic families and 112 endemic genera of plants, and 1 endemic family and 3 endemic genera of birds. In contrast, the United States and Canada, with an expanse 8.8 times larger than the 25 hotspots combined, have only two endemic families of plants. Moreover, plant family richness can often serve as a predictor of species richness for certain animal taxa such as mammals, amphibians and reptiles³².

Action responses

In sum, the 25 hotspots contain the sole remaining habitats of 44% of the Earth's plant species and 35% of its vertebrate species, and these habitats face a high risk of elimination. Many of the hotspots could well contain sizeable proportions of endemic invertebrates. It is often supposed¹⁻⁴ that, were the present mass extinction of species to proceed virtually unchecked, between one-third and two-thirds of all species would be likely to disappear within the foreseeable future. The hotspots analysis indicates that much of this problem could be countered through protection of the 25 hotspots.

An aggregate expanse of 800,767 square kilometres, 38% of the hotspots total, is already protected in parks and reserves. True, some of these are little better than 'paper parks', but they offer a modicum of legal status. All are in urgent need of stronger safeguards, including those five hotspots where the protected expanse is as large as the hotspot itself. The areas without any protection at all

Table 4 Species/area ratios per 100 km² of hotspots

Hotspot	Endemic plants	Endemic vertebrates
Tropical Andes	6.4	0.5
Mesoamerica	2.2	0.5
Caribbean	23.5	2.6
Brazil's Atlantic Forest	8.7	0.6
Choco/Darien/Western Ecuador	3.6	0.7
Brazil's Cerrado	1.2	0.03
Central Chile	1.8	0.06
California Floristic Province	2.7	0.09
Madagascar	16.4	1.3
Eastern Arc and Coastal Forests of Tanzania/Kenya	75	6.1
Western African Forests	1.8	0.2
Cape Floristic Province	31.6	0.3
Succulent Karoo	6.5	0.15
Mediterranean Basin	11.8	0.2
Caucasus	3.2	0.1
Sundaland	12.0	0.6
Wallacea	2.9	1.0
Philippines	64.7	5.7
Indo-Burma	7.0	0.5
South-Central China	5.5	0.3
Western Ghats/Sri Lanka	17.5	2.9
SW Australia	13.0	0.3
New Caledonia	49.1	1.6
New Zealand	3.1	0.2
Polynesia/Micronesia	33.3	2.2

Table 5 Congruence between plants and vertebrates

Hotspot	Endemic plants as % of global total (300,000)	Endemic vertebrates as % of global total (27,298)	Congruence (%) (rounded)
Tropical Andes	6.7%	5.7%	85
Mesoamerica	1.7%	4.2%	41
Caribbean	2.3%	2.9%	79
Brazil's Atlantic Forest	2.7%	2.1%	78
Choco/Darien/Western Ecuador	0.8%	1.5%	53
Brazil's Cerrado	1.5%	0.4%	27
Central Chile	0.5%	0.2%	40
California Floristic Province	0.7%	0.3%	43
Madagascar	3.2%	2.8%	88
Eastern Arc and Coastal Forests of Tanzania/Kenya	0.5%	0.4%	80
West African Forests	0.8%	1.0%	80
Cape Floristic Province	1.9%	0.2%	11
Succulent Karoo	0.6%	0.2%	33
Mediterranean Basin	4.3%	0.9%	21
Caucasus	0.5%	0.2%	40
Sundaland	5.0%	2.6%	52
Wallacea	0.5%	1.9%	26
Philippines	1.9%	1.9%	100
Indo-Burma	2.3%	1.9%	83
South-Central China	1.2%	0.7%	58
Western Ghats/Sri Lanka	0.7%	1.3%	54
SW Australia	1.4%	0.4%	29
New Caledonia	0.9%	0.3%	33
New Zealand	0.6%	0.5%	83
Polynesia/Micronesia	1.1%	0.8%	73

amount to 1.3 million square kilometres or 62% of the total area of the hotspots. This expanse surely represents the greatest biodiversity challenge of the foreseeable future, and should be safeguarded through, for example, a 'hotspots rescue fund'. In some areas, outright protection is still the best option. In other areas, this is not feasible because of human settlements and other activities long in place. These areas could receive a measure of protection as 'conservation units' that allow some degree of multiple use provided that species safeguards are always paramount.

This is not to say that protection of the hotspots would safeguard all their species indefinitely. According to the well-established theory of island biogeography³³, when an area loses a large proportion of its original habitat and especially when the remaining habitat is severely fragmented, it will eventually lose some of its species through what are technically known as 'ecological equilibration' or delayed fallout effects. There is much empirical evidence to support this; for instance, the loss of birds in Brazil's Atlantic forest³⁴, in Southeast Asia's forests³⁵, in tropical forests generally^{36,37} and in the United Kingdom³⁸; of tree species in tropical forests³⁹; of forest plants in eastern North America⁴⁰; of primates in Africa's forests⁴¹; of large mammals in Tanzania⁴²; and of species generally⁴³.

Consider the consequences for the smallest hotspot, the Eastern Arc. The remaining primary vegetation is only 6.7% of the original, and its expanse of 2,000 km² is split into no fewer than 128 patches ranging in size from over 100 to 10 or fewer square kilometres. A bigger hotspot, Cape Floristic Province, with an expanse of 18,000 km² and 24.3% of its original primary vegetation, is spread

around several thousand patches ranging from over 100 to 0.1 km².

Although most island-biogeography losses are not likely to ensue for some time, it makes sense to take immediate steps to safeguard the hotspots to avoid an exceptionally large extinction spasm through outright loss of habitat on a scale to swamp island biogeography impacts. As for past extinctions in the hotspots, all too little is known with respect to taxa across the board including invertebrates; however, if we use birds extinct since 1800 as a surrogate we find that nearly 80% of those that disappeared were from hotspot areas.

These considerations apart, the prospect of a mass extinction can be made far less daunting and much more manageable through the hotspots strategy, with its tight targeting of conservation efforts.

The hotspots findings accord well with several other priority-setting analyses. There is a 68% overlap with Birdlife International's Endemic Bird Areas⁴⁴, 82% with IUCN/WWF International's Centres of Plant Diversity and Endemism¹⁷ and 92% with the most critical and endangered eco-regions of WWF/US's Global 200 List⁴⁵. The hotspots approach is more comprehensive than the first two because it combines five categories of species, and it is more closely focused than the third.

Other areas appear to feature exceptional plant endemism and exceptional threat, but are not sufficiently documented to meet the hotspots criteria. They include the Ethiopian Highlands, the Angola Escarpment, southeastern China, Taiwan, and the forests of the Albertine Rift in eastern Democratic Republic of Congo (formerly Zaire), southwestern Uganda and northern Rwanda. Much better known and with a high species/area ratio but without sufficient endemic plant species to qualify as a hotspot is the so-called Wet Tropics and adjacent tropical forest tracts along the Queensland coast in Australia, containing around 1,200 endemic plants in less than 11,000 km². Adding these areas to the hotspots list would increase the total of plants endemics by only a few per cent.

In addition, there are a few tropical forest expanses, known as 'major wilderness areas'⁴⁶ or 'good news' areas^{5,6}. They total some 6–7 million km² and feature concentrations of endemic species while retaining at least 75% of their primary vegetation, and have fewer than five people per square kilometre. One is the island of New Guinea, with around 15,000 endemic plants. Others include the Guayana Shield of northeastern Amazonia, the lowlands of western Amazonia and the Congolian Forest, with perhaps another 30,000 endemic plants. Were these regions to compose a supplementary conservation strategy, they could increase the number of plants endemics to almost 60% of all plant species in roughly 5% of the Earth's land surface.

Funding

Since the original hotspots strategy^{5,6} began to be implemented in 1989, some \$400 million has been invested by the MacArthur Foundation, the W. Alton Jones Foundation, Conservation International, the World Wildlife Fund and other non-governmental organizations. An annual average of \$40 million over 10 years is only a tiny fraction of the amount spent per year on biodiversity

Table 6 The eight hottest hotspots in terms of five factors

Hotspot	Endemic plants		Endemic vertebrates		Endemic plants/ area ratio (species per 100 km ²)		Endemic vertebrates/area ratio (species per 100 km ²)		Remaining primary vegetation as % of original extent	Times appearing in top 10 for each of five factors
Madagascar	9,704	4	771	4	16.4	8	1.3	7	9.9	5
Philippines	5,832	8	518	9	64.7	2	5.7	2	3.0	5
Sundaland	15,000	2	701	5	12.0	10	0.6	10=	7.8	5
Brazil's Atlantic Forest	8,000	5	654	6	8.7		0.6	10=	7.5	4
Caribbean	7,000	6=	779	3	23.5	6	2.6	4	11.3	4
Indo-Burma	7,000	6=	528	8	7.0		0.5		4.9	3
Western Ghats/Sri Lanka	2,180		355		17.5	7	2.9	3	6.8	3
Eastern Arc and Coastal Forests of Tanzania/Kenya	1,500		121		75.0	1	6.1	1	6.7	3

conservation by governments and international agencies, these funds being assigned mainly to across-the-board activities rather than the concentrated efforts advocated here. The traditional scattergun approach of much conservation activity, seeking to be many things to many threatened species, needs to be complemented by a 'silver bullet' strategy in the form of hotspots with their emphasis on cost-effective measures.

We could go far towards safeguarding the hotspots and thus a large proportion of all species at risk for an average of \$20 million per hotspot per year over the next five years, or \$500 million annually. Although this is 12.5 times the annual average of the \$400 million spent on hotspots over the past decade, it is still only twice the cost of a single Pathfinder mission to Mars, which has been justified largely on biodiversity grounds (the search for extraterrestrial life). The \$500 million annually is to be compared, moreover, with a recent estimate⁴⁷ for a comprehensive conservation programme to protect biodiversity world-wide costing \$300 billion annually—a total that should, in turn, be compared with subsidies of various sorts that degrade environments and economies alike, amounting to \$1.5 trillion annually world-wide⁴⁸.

Finally, recall that the mass extinction of species, if allowed to persist, would constitute a problem with far more enduring impact than any other environmental problem. According to evidence from mass extinctions in the prehistoric past, evolutionary processes would not generate a replacement stock of species within less than several million years. What we do (or do not do) within the next few decades will determine the long-term future of a vital feature of the biosphere, its abundance and diversity of species. This expanded hotspots strategy offers a large step toward avoiding an impoverishment of the Earth lasting many times longer than *Homo sapiens* has been a species.

Received 22 September; accepted 22 December 1999.

- Ehrlich, P. R. Energy use and biodiversity loss. *Phil. Trans. R. Soc. Lond. B* **344**, 99–104 (1994).
- Myers, N. Two key challenges for biodiversity: discontinuities and synergisms. *Biodiversity Cons.* **5**, 1025–1034 (1996).
- Pimm, S. L., Russell, G. J., Gittleman, J. L. & Brooks, T. M. The future of biodiversity. *Science* **269**, 347–350 (1995).
- Wilson, E. O. *The Diversity of Life* (Belknap, Cambridge, Massachusetts, 1992).
- Myers, N. Threatened biotas: 'hotspots' in tropical forests. *Environmentalist* **8**, 187–208 (1988).
- Myers, N. The biodiversity challenge: expanded hotspots analysis. *Environmentalist* **10**, 243–256 (1990).
- Pressey, R. L., Humphries, C. J., Margules, C. R., Vane-Wright, R. I. & Williams, P. H. Beyond opportunism: key principles for systematic reserve selection. *Trends Ecol. Evol.* **8**, 124–128 (1993).
- Prendergast, J. R., Quinn, R. M. & Lawton, J. H. The gaps between theory and practice in selecting nature reserves. *Cons. Biol.* **13**, 484–492 (1999).
- Ginsberg, J. Global conservation priorities. *Cons. Biol.* **13**, 5 (1999).
- Dobson, A. P., Rodriguez, J. P., Roberts, W. M. & Wilcove, D. S. Geographic distribution of endangered species in the United States. *Science* **275**, 550–553 (1997).
- Reid, W. V. Biodiversity hotspots. *Trends Ecol. Evol.* **13**, 275–280 (1998).
- Prendergast, J. R., Quinn, R. M., Lawton, J. H., Eversham, B. C. & Gibbons, D. W. Rare species, the coincidence of diversity hotspots and conservation strategies. *Nature* **365**, 335–337 (1993).
- Williams, P. et al. A comparison of richness hotspots, rarity hotspots, and complementary areas for conserving diversity of British birds. *Cons. Biol.* **10**, 155–174 (1996).
- Vane-Wright, R. I., Humphries, C. J. & Williams, P. H. What to protect?—systematics and the agony of choice. *Biol. Cons.* **55**, 235–254 (1991).
- Williams, P. H., Humphries, C. J. & Vane-Wright, R. I. Measuring biodiversity: taxonomic relatedness for conservation priorities. *Aust. Syst. Bot.* **4**, 665–679 (1991).
- Mittermeier, R. A., Myers, N., Gil, P. R. & Mittermeier, C. G. *Hotspots: Earth's Biologically Richest and Most Endangered Terrestrial Ecoregions* (Cemex, Conservation International and Agrupacion Sierra Madre, Monterrey, Mexico, 1999).

- Davis, S., Heywood, V. H. & Hamilton, A. C. (eds) *Centres of Plant Diversity* (three vols) (World Wide Fund for Nature and International Union for Conservation of Nature and Natural Resources, Gland, Switzerland, 1994–1997).
- Groombridge, B. (ed.) *Global Biodiversity* (Chapman and Hall, London, 1992).
- Heywood, V. H. (ed.) *Global Biodiversity Assessment* (Cambridge Univ. Press, Cambridge, 1995).
- Prance, G. T., Beent J. H., Dransfield, J. & Johns, R. *The Tropical Flora Remains Undercollected* (Missouri Botanical Garden Scientific Publications, St. Louis, Missouri, in the press).
- Nowak, R. *Walker's Mammals of the World* (Johns Hopkins Univ. Press, Baltimore, Maryland, 1999).
- Sibley, C. G. & Monroe, B. L. *Distribution and Taxonomy of Birds of the World* (Yale Univ. Press, New Haven, Connecticut, 1990).
- Uetz, P. & Etzold, T. The EMBL/EBI reptile database. *Herpetol. Rev.* **27**, 175 (1996).
- Glaw, F. & Kohler, J. Amphibian species diversity exceeds that of mammals. *Herpetol. Rev.* **29**, 11–12 (1998).
- Eschmeyer, W. M. *Catalog of Fishes* (California Academy of Sciences, San Francisco, 1998).
- Janzen, D. H. How to be a fig. *Annu. Rev. Ecol. Systemat.* **10**, 13–51 (1979).
- Farrell, B. D. 'Inordinate Fondness' explained: why are there so many beetles? *Science* **281**, 555–557 (1998).
- Gaston, K. J. Regional numbers of insect and plant species. *Funct. Ecol.* **6**, 243–247 (1991).
- Strong, D. R., Lawton, J. H. & Southwood, T. R. E. *Insects on Plants: Community Patterns and Mechanisms* (Blackwell, Oxford, 1984).
- Price, P. W. *Insect Ecology* 3rd edition (Wiley, New York, 1997).
- Balmford, A. & Long, A. Across-country analyses of biodiversity congruence with current conservation efforts in the tropics. *Cons. Biol.* **9**, 1539–1547 (1996).
- Williams, P. H., Gaston, K. & Humphries, C. J. Mapping biodiversity value worldwide: combining higher-taxon richness from different groups. *Proc. R. Soc. Lond. B* **264**, 141–148 (1997).
- MacArthur, R. H. & Wilson, E. O. *The Theory of Island Biogeography* (Princeton Univ. Press, Princeton, 1967).
- Brooks, T. & Balmford, A. Atlantic forest extinctions. *Nature* **380**, 115 (1996).
- Brooks, T., Pimm, S. L. & Collar, N. J. Deforestation predicts the number of threatened birds in Insular Southeast Asia. *Cons. Biol.* **11**, 382–394 (1997).
- Brooks, T. M., Pimm, S. L. & Oyugi, J. O. Time lag between deforestation and bird extinction in tropical forest fragments. *Cons. Biol.* **13**, 1140–1150 (1999).
- Laurance, W. F. Introduction and synthesis. *Biol. Cons.* **91**, 101–107 (1999).
- Gaston, K. J. & Nicholls, A. O. Probable times to extinction of some rare breeding bird species in the United Kingdom. *Proc. R. Soc. Lond. B* **259**, 119–123 (1995).
- Turner, I. M. Species loss in fragments of tropical rain forests: a review of the evidence. *J. Appl. Ecol.* **33**, 200–209 (1996).
- Pimm, S. L. & Askins, R. A. Forest losses predict bird extinctions in Eastern North America. *Proc. Natl Acad. Sci. USA* **92**, 9343–9347 (1995).
- Cowlinshaw, G. Predicting the pattern of decline of African primate diversity: an extinction debt from historical deforestation. *Cons. Biol.* **13**, 1183–1193 (1999).
- Newmark, W. D. Insularization of Tanzanian parks and the local extinction of large mammals. *Cons. Biol.* **10**, 1549–1556 (1996).
- Tilman, D., May, R. M., Lehman, C. L. & Nowak, M. A. Habitat destruction and the extinction debt. *Nature* **371**, 65–66 (1994).
- Stattersfield, A. J., Crosby, M. J., Long, A. J. & Wege, D. C. *Endemic Bird Areas of the World: Priorities for Biodiversity Conservation* (Birdlife International, Cambridge, UK, 1998).
- Dinerstein, E. et al. *The Global 200: Key Ecoregions for Saving Life on Earth* (World Wildlife Fund-US, Washington DC, 1996).
- Mittermeier, R. A., Myers, N., Thomsen, J. B., da Fonseca, G. A. B. & Olivieri, S. Biodiversity hotspots and major tropical wilderness areas: approaches to setting conservation priorities. *Cons. Biol.* **12**, 516–520 (1998).
- James, A. N., Gaston, K. J. & Balmford, A. Balancing the Earth's accounts. *Nature* **401**, 323–324 (1999).
- Myers, N. Lifting the veil on perverse subsidies. *Nature* **392**, 327–328 (1999).

Supplementary Information is available from Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of *Nature*. This and further information is also available at <http://www.conservation.org>.

Acknowledgements

We thank P. Robles Gil of Agrupacion Sierra Madre and the scientists listed in Supplementary Information for their help with information and analysis; P. Chambers, S. Norris and M. Prescott for research help; and D. Duthie and J. McNeely for comments on an early draft. We also thank the Mexican company CEMEX for its major financial support, and the MacArthur Foundation and S. Concannon for additional support.

Correspondence and requests for materials should be addressed to N.M. (e-mail: myers1n@aol.com).

Structural basis for recognition and repair of the endogenous mutagen 8-oxoguanine in DNA

Steven D. Bruner, Derek P. G. Norman & Gregory L. Verdine

Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, Massachusetts 02138, USA

Spontaneous oxidation of guanine residues in DNA generates 8-oxoguanine (oxoG). By mispairing with adenine during replication, oxoG gives rise to a G-C → T-A transversion, a frequent somatic mutation in human cancers. The dedicated repair pathway for oxoG centres on 8-oxoguanine DNA glycosylase (hOGG1), an enzyme that recognizes oxoG-C base pairs, catalysing expulsion of the oxoG and cleavage of the DNA backbone. Here we report the X-ray structure of the catalytic core of hOGG1 bound to oxoG-C-containing DNA at 2.1 Å resolution. The structure reveals the mechanistic basis for the recognition and catalytic excision of DNA damage by hOGG1 and by other members of the enzyme superfamily to which it belongs. The structure also provides a rationale for the biochemical effects of inactivating mutations and polymorphisms in hOGG1. One known mutation, R154H, converts hOGG1 to a pro-mutator by relaxing the specificity of the enzyme for the base opposite oxoG.

Nearly all higher organisms generate energy by aerobic respiration, a process that involves the stepwise four-electron reduction of molecular oxygen to water. The partially reduced species that are produced as intermediates and by-products of aerobic respiration, including $O_2^{\cdot-}$, H_2O_2 and $OH^{\cdot}$, are potent electrophilic oxidants that can escape the mitochondria and attack vital cellular components. These same oxidants, collectively called reactive oxygen species (ROS), are also generated in cells through exposure to ionizing radiation and other agents that generate free radicals. A major target of ROS and free-radical attack is the cellular genome, which suffers the formation of numerous genotoxic adducts and DNA strand breaks¹. As these lesions are believed to have a key role in pathophysiologic processes such as cancer and ageing, much attention has been focused on understanding the mechanisms of oxidative DNA damage and repair².

Among the most deleterious of ROS-induced adducts is 7,8-dihydro-8-oxoguanine (oxoG, Fig. 1), which is formed by the oxidation of the guanine base in DNA^{3,4}. Whereas guanine bases

have a strong preference to pair with cytosine, oxoG can pair both in Watson–Crick mode with cytosine and in Hoogsteen mode with adenine. In the latter pairing mode, oxoG residues in DNA frequently mispair with adenine during replication *in vitro* and *in vivo*, ultimately giving rise to G-C → T-A transversion mutations (Fig. 1a). These transversions are the second most common somatic mutation found in human cancers and are especially prevalent in the mutational spectrum of the tumour suppressor gene *p53* (ref. 5), facts that have implicated oxoG as a major endogenous mutagen that contributes broadly to spontaneous cell transformation.

The cellular defence system against oxoG mutagenesis^{3,4} has been extensively studied in bacteria, and progress has been made on cloning and characterizing the functionally related system in eukaryotes². One component is MutT (hMTH1 in humans), a triphosphatase that cleanses the nucleotide precursor pool by catalysing the hydrolysis of oxo-dGTP to oxo-dGMP and inorganic pyrophosphate. The second component is MutY (hMYH1 in humans), a DNA glycosylase that initiates the repair of misreplicated

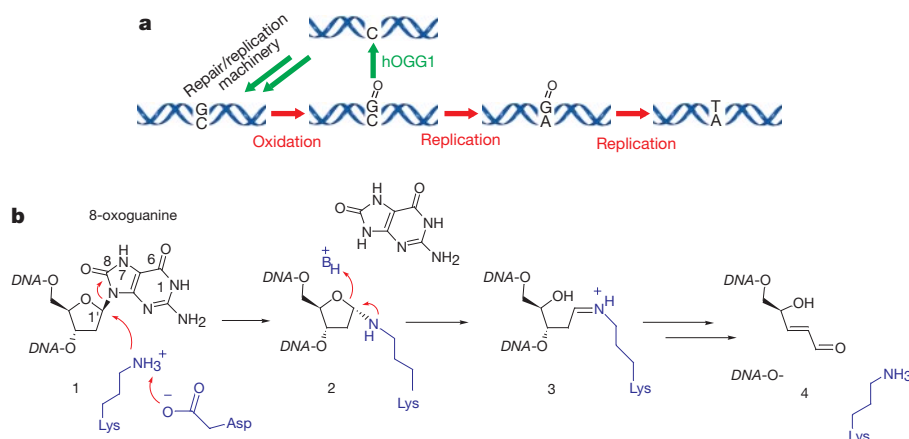


Figure 1 The role of hOGG1 in preventing mutagenesis by oxidized guanines in DNA. **a**, Direct oxidation of guanine residues in DNA generates 8-oxoguanine opposite cytosine (oxoG-C). hOGG1 can recognize the lesion and excise the oxoG base, leaving behind an abasic site that is restored to guanine by the cellular abasic site repair/replication machinery. If the oxoG strand is replicated instead, the lesion preferentially mispairs with adenine to give an oxoG-A intermediate. Replication of the A strand in the oxoG-A

intermediate completes the mutational process, resulting in an overall G-C → T-A transversion mutation. **b**, Enzymatic mechanism of hOGG1. The enzyme uses an active-site lysine residue (Lys 249) to attack C-1' of the deoxyribose sugar and expel the oxoG base. Further enzymatic steps lead to strand cleavage at the 3'-phosphate.

oxoG-A pair by catalysing hydrolysis of the glycosidic bond linking the adenine base to its sugar. The third and final component is Ogg1, a DNA glycosylase/ β -lyase that recognizes oxoG opposite cytosine. Ogg1 orthologues have subsequently been cloned from mice and humans^{6–13}. Inactivation of the *OGG1* gene in yeast generates a mutator phenotype that is specific for G-C $\rightarrow$ T-A transversions^{14,15}. Mice with targeted disruption of the murine *OGG1* gene appear to develop normally, but have an elevated rate of spontaneous mutagenesis in nonproliferative tissues and accumulate abnormally high levels of oxoG in their genome¹⁶.

Humans express two main splice-variant forms of the OGG1 protein, which are identical in their first 316 amino acids but differ completely at the carboxy terminus. Signal sequences present in the divergent C-terminal segments target the 345-amino-acid (relative molecular mass, 39K) α -hOGG1 form to the nucleus and the 424-amino-acid (47K) β -hOGG1 form to the mitochondrion^{17,18}. Both forms of hOGG1 exhibit the same catalytic activity, that is, the ability to excise oxoG and the structurally related lesion 2,6-diamino,4-hydroxy-5-formamidopyrimidine (FaPy, a hydrolytic ring-opening product of guanine) (glycosylase activity) and to cleave the DNA backbone (β -lyase activity)^{6,8–11,13,19,20}.

The human OGG1 gene is located in a region on the short arm of chromosome 3 (3p26) that is subject to monoallelic deletion resulting in loss of heterozygosity in nearly 100% of small-cell lung cancers and more than 50% of renal, salivary and non-small-cell lung cancers, and in a smaller but significant percentage of cancers of the breast, head and neck. Loss of one hOGG1 allele is expected to increase moderately the mutagenic burden imposed by guanine oxidation. However, loss of the remaining allele would abolish hOGG1 activity altogether, thereby placing the cell at great risk, especially under conditions of oxidative stress that is induced,

for example, by exposure to cigarette smoke. Recognition of these facts has fuelled efforts to identify mutations in *hOGG1* and to determine their effects on the biochemical activity of the protein^{21–24}. Together, these studies provide tantalizing evidence that inactivating mutations in the *hOGG1* gene contribute to a small but significant proportion of human cancers.

Here we report the structure of the core catalytic domain of hOGG1 bound to an oxoG-C-containing duplex oligonucleotide at 2.1 Å resolution. The structure clearly reveals the basis for recognition of the oxoG lesion and its complementary cytosine, and enables the biochemical effects of somatic mutations to be rationalized. We find that one such mutation actually alters the specificity of the protein so as to favour promutagenic repair of oxoG opposite bases other than cytosine. The structural results presented here provide insight into the ways in which the major superfamily of base-excision DNA repair proteins recognize and process their substrates.

Structure determination

A mutant form of hOGG1 in which Lys 249 is changed to glutamine has been shown to lack catalytic activity but retain substrate recognition (see Fig. 1)²⁵. Crystallization trials with full-length K249Q hOGG1 and a panel of oxoG-C-containing oligonucleotides failed to yield high-quality crystals. We used limited tryptic digestion of the protein to remove unstructured peptide segments at the amino- and carboxy-terminal ends of the protein, thus yielding a core domain comprising residues 12–327, which was then over-expressed and purified. The wild-type core domain exhibits the same ability to discriminate for oxoG opposite cytosine as does full-length hOGG1, but binds DNA with slightly lower affinity (data not shown), presumably owing to removal of the polycationic

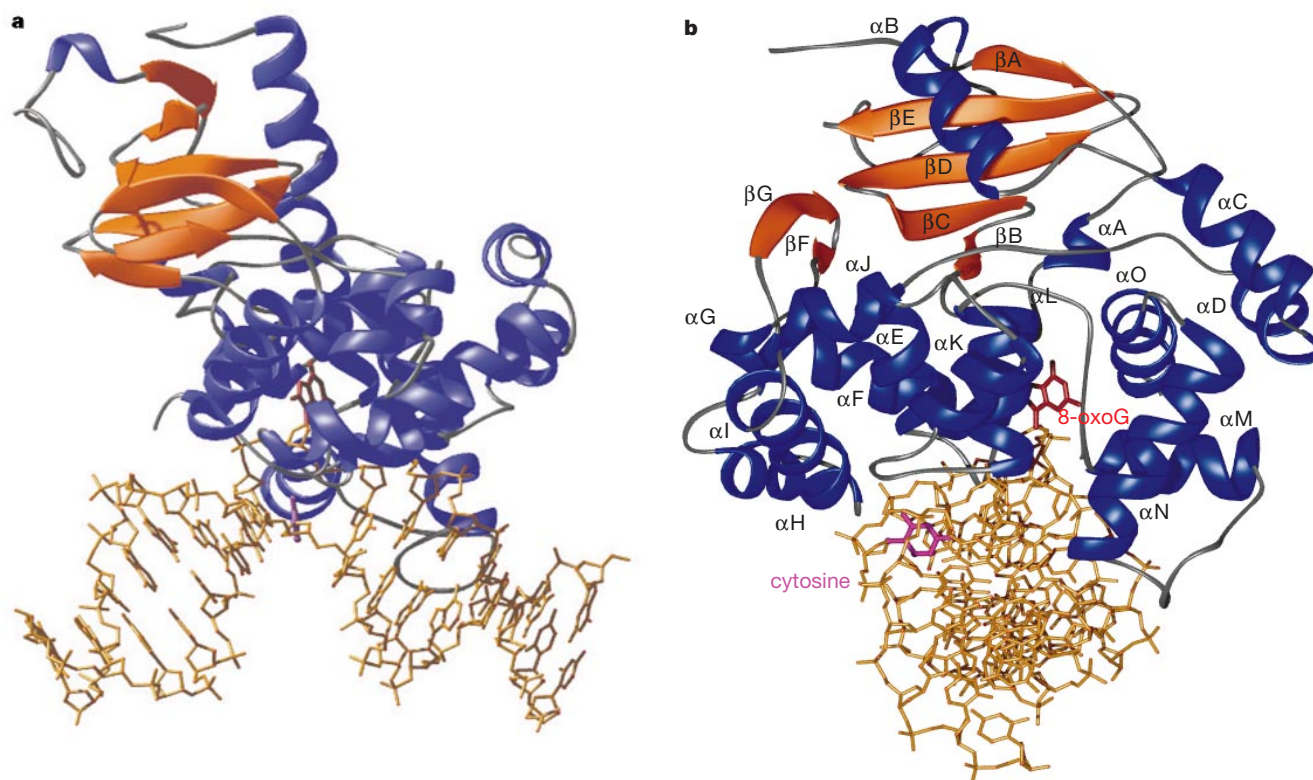


Figure 2 Overall structure of the hOGG1–DNA complex. **a, b**, Two orthogonal views of the hOGG1–DNA complex with the protein shown in ribbon representation (blue, α -helices; orange, β -sheets; grey, non-repetitive elements) and 15-base DNA oligonucleotide in stick representation (gold). The substrate oxoG base (red) is completely extruded from the DNA helix and inserted into an active-site pocket. The complementary or ‘estranged’

cytosine (purple) remains stacked in the helix. The enzyme bends the DNA ($\sim 70^\circ$) at the plane of the oxoG-C base pair. The bend in the DNA exposes the edge of the estranged cytosine to protein side chains that make specific contacts (see text). Figure generated using RIBBONS⁴³ and rendered with POV-ray.

nuclear localization signal sequence at the C terminus (residues 335–342).

In complex with an oxoG-C-containing duplex 15-base oligomer, the recombinant core domain (K249Q mutant) yielded co-crystals suitable for structure determination by multiwavelength anomalous diffraction (MAD) phasing (Table 1). The final model contains residues 12–325, and excludes three residues (80–82) that are located on a disordered loop.

General features of the hOGG1–DNA complex

The overall structure of the hOGG1–DNA complex is shown in Fig. 2a and b. As predicted, the protein contains a fold that is common to members of a superfamily of base-excision DNA repair proteins¹⁹, typified by *Escherichia coli* endonuclease III and AlkA. Members of this superfamily are found in organisms ranging from bacteria to mammals, and repair a wide variety of base lesions resulting from, for example, DNA oxidation, alkylation, hydration and deamination. The hallmark of these proteins is a helix–hairpin–helix²⁶ structural element followed by a Gly/Pro-rich loop and a conserved aspartic acid (HhH–GPD motif)^{19,27} (Fig. 3). Although several structures of HhH–GPD superfamily members have been reported^{27–29}, none of these contained a bound DNA substrate. The structure of hOGG1 contains, in addition to the two α -helical domains common to all known family members, a third antiparallel β -sheet domain found thus far in only the alkylation repair DNA

glycosylase AlkA^{27,30} (Fig. 3). Although hOGG1 and AlkA clearly have the same overall fold, the two structures differ in many respects at a detailed level (overall r.m.s. deviation, 3.1 Å).

The protein is firmly latched onto the DNA backbone of the oxoG-containing strand, but does not contact the backbone of the complementary strand at all (Fig. 2a,b). The substrate oxoG residue is fully extruded from the helix and is inserted deeply into an extrahelical active-site cleft on the enzyme. The complementary cytosine is intrahelical, but is largely unstacked from its neighbouring bases owing to a pronounced kink in the DNA, which also gives rise to sharp ($\sim 70^\circ$) bending of the duplex away from hOGG1. In the crystal lattice, one end of the duplex is coaxially stacked with a neighbouring complex whereas the other end is free in solution, making it unlikely that crystal packing influences bending. Outside the active site, the conformation of the DNA is nearly canonical B-form.

Protein–DNA recognition

The marked remodelling of DNA structure caused by hOGG1 is induced by extensive interactions between the protein and the phosphates of the oxoG-containing strand, the substrate oxoG residue and the complementary cytosine (see Fig. 4a), which altogether comprise 2,268 Å² of contact surface area. Whereas most DNA-binding proteins use surfaces that are rich in lysine and arginine residues to contact backbone phosphates, hOGG1 binds the DNA backbone in a channel that is nearly charge neutral (Fig. 5) with only a single basic residue, His 270. A striking feature of the complex is the large number of the α -helices in hOGG1 that have their N-terminal ends orientated toward the DNA (Fig. 2a,b). This disposition of α -helices maximizes the helix–dipole interaction, suggesting that hOGG1 may use dipolar electrostatic interactions more so than the usual salt bridges to bind its DNA substrate. Only one of these helices, α L, actually contacts the DNA backbone. This is noteworthy because α L and the loop and helix (α K) preceding it comprise the conserved helix–hairpin–helix (HhH) element. In addition to the phosphate contacts made by the main chain of α L to p^{−2} (Val 250 and Gln 249), a highly conserved glycine residue of the loop (Gly 245) hydrogen bonds to p^{−3}. The HhH contacts a region of the DNA substrate on the 3' side of the lesion that has a nearly normal B-form conformation; hence the HhH appears primarily to be involved in positioning the duplex for presentation of the lesion into the active-site cavity.

Phosphate backbone interactions with p^{−1}, p⁰ and p¹ are important in stabilizing the highly unusual DNA backbone conformation at the site of the lesion. Bond rotations necessary to swivel the oxoG residue out of the helix cause the non-bridging oxygens of p^{−1} to point inward toward the helix axis. Because of this inward orientation and bending of the helix, p^{−1} ends up lying within 7 Å of p¹, half the corresponding distance in normal B-form DNA. To relieve the electrostatic repulsion caused by the proximity of p^{−1} and p¹, a partially hydrated Ca²⁺ ion derived from the crystallization media binds between them, coordinating directly to p¹ and through a bridging water to p^{−1}. Although the Ca²⁺ ion, which may be replaced by Mg²⁺ under physiological conditions, does not interact directly or indirectly with the protein, its H₂O ligands form hydrogen bonds with the DNA, thus stabilizing its bent and kinked conformation. The central residue (Asn 150) of a conserved NNN motif in hOGG1 (Fig. 4b) also stabilizes the inward orientation of p^{−1} through hydrogen bonding. On the 5' side of the oxoG residue, His 270 forms a hydrogen bond with p⁰, but the protein makes no further backbone interactions.

The oxoG-specificity pocket

The structure of the hOGG1–DNA complex provides a clear solution to the long-standing puzzle of how the cellular repair machinery recognizes oxoG in DNA amidst the vast excess of guanine, a close structural relative. The oxoG residue is fully

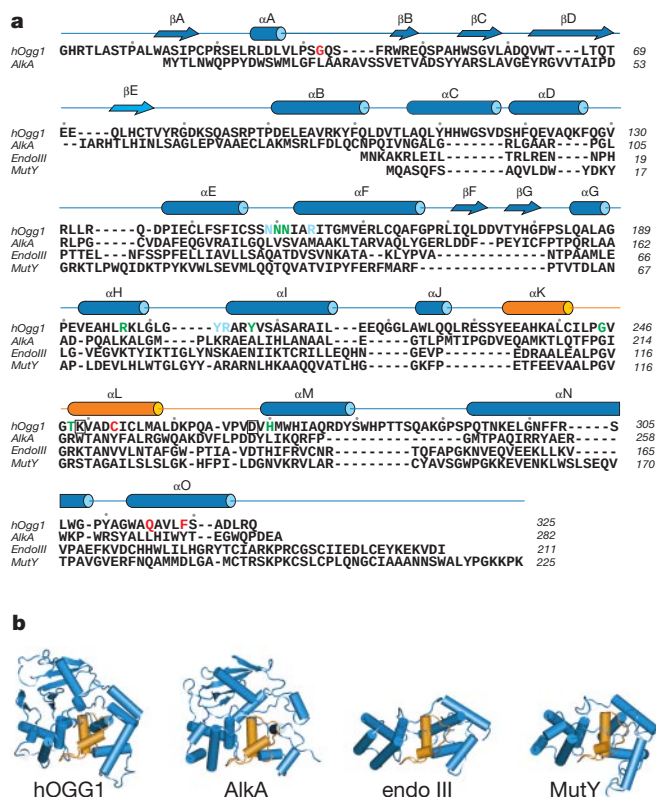


Figure 3 Sequence/structural alignment of hOGG1 with other HhH–GPD superfamily members. **a**, Structure-based alignment of the hOGG1 sequence with those of *E. coli* AlkA^{27,30}, *E. coli* endonuclease III (refs 26, 28) and *E. coli* MutY²⁹. Secondary structure assignments are listed above the primary sequence with α -helices highlighted by cylinders and β -sheets highlighted as arrows. The highly conserved HhH–GPD motif is shown in orange. Residues in hOGG1 are highlighted as follows: the catalytic Lys 249 and Asp 268 are boxed; residues that interact with the oxoG and estranged cytosine are red and blue, respectively; residues making DNA backbone contacts are green. **b**, The conserved HhH–GPD motif (orange) in structurally characterized members of the HhH–GPD superfamily.

extruded from the DNA helix and inserted into an extrahelical active-site pocket on the enzyme, consistent with structure-based predictions made for other members of the HhH-GPD superfamily^{26–29}. Although the 8-oxo substituent in oxoG causes this nucleoside to favour the *syn* conformer about the glycosidic bond, oxoG binds the active site of hOGG1 in an *anti* conformation quite similar to that in normal duplex DNA. The extruded DNA backbone and glycosyl conformations combine to extend the oxoG far out from the DNA helix, thus enabling the nucleoside to plunge deeply into the active-site pocket of hOGG1. The space in the duplex vacated by the oxoG residue is filled by a residue of the NNN motif, Asn 149, which forms a hydrogen bond through its side-chain amide carbonyl with the exocyclic NH₂ of the ‘estranged’ cytosine (C⁰).

The oxoG-specificity pocket on the enzyme envelops the lesion base, interacting with every accessible surface except the minor-groove edge (Fig. 6a). The most characteristic feature of oxoG—its 8-oxo-carbonyl function—is completely devoid of any interacting

partner. Instead, the enzyme recognizes the urea system in oxoG by virtue of its N7-H, which donates a hydrogen bond to the main-chain carbonyl of Gly 42. Because the N7 atom of guanine presents an acceptor lone pair instead of a donor proton, its interaction with Gly 42 is expected to be repulsive. Indeed, of all the contacts made to the oxoG base, that involving Gly 42 is the only one that would clearly be different with oxoG versus guanine; thus, the responsibility for discriminating oxoG from guanine appears to be borne by a single hydrogen bond. It is noteworthy that Gly 42, despite its critical importance, is the only residue that the β -sheet domain contributes to the hOGG1–DNA interface.

The other residues in the hOGG1 oxoG-specificity pocket appear exquisitely capable of recognizing the lesion while excluding A, C and T. Phe 319 and Cys 253 interact with opposite π -faces of the oxoG, sandwiching the base in the active site. The Gln 315 amide-NH₂ and a tightly bound water molecule cooperate to recognize O⁶, and the side-chain carbonyl of Gln 315 forms hydrogen bonds with both N1 and N²H. A second tightly bound water is hydrogen bonded to O⁶. Gln 315, the captive waters and Gly 42 are not chemically matched to the hydrogen bond donor/acceptor pattern on A, C and T.

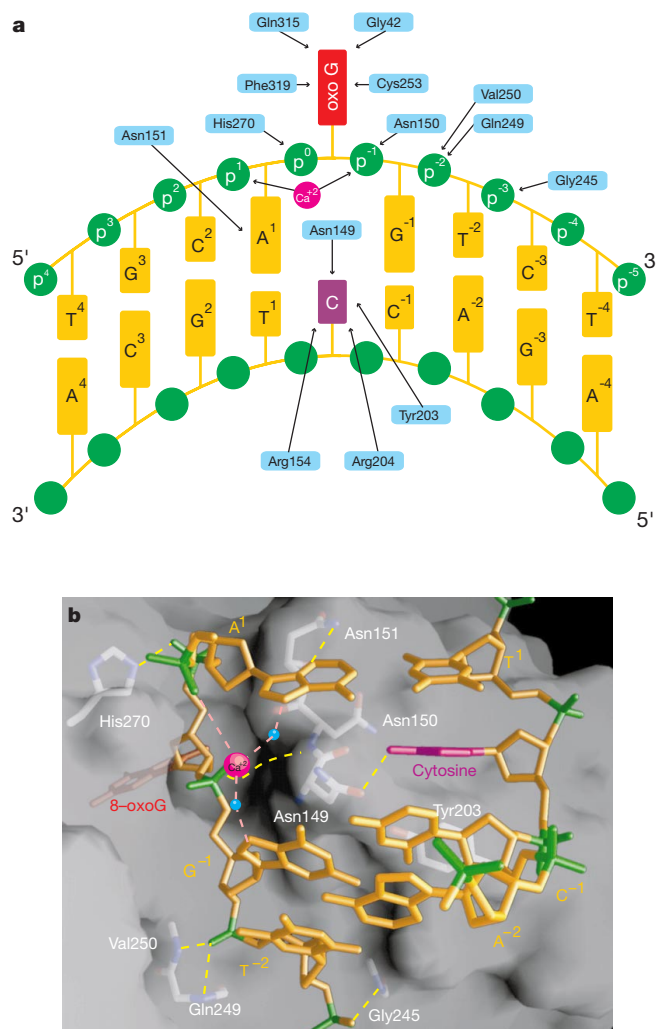


Figure 4 Contact interface between hOGG1 and oxoG-containing DNA. **a**, Diagram of the DNA–protein interface. Arrows indicate interactions between amino-acid residues (blue) and oxoG (red), estranged cytosine (purple), backbone phosphates (green) and DNA bases (yellow). **b**, Surface diagram (prepared with GRASP⁴⁴) of the DNA–protein interface. Hydrogen bonds are shown in yellow and coordinative interactions with a calcium ion (magenta) and its associated waters (blue) are shown in pink. The wedging of Tyr 203 into the helix below the estranged cytosine helps to create the sharp bend and kink in the DNA at that site. The DNA bends away from the residues of the conserved NNN motif, Asn 151, Asn 150 and Asn 149, which contacts DNA from behind (minor-groove face) in this view.

Specific recognition of the estranged cytosine

As described above, the extrusion of oxoG leaves its partner cytosine (C⁰) estranged in the helix. The enzyme resorts to extraordinary means to recognize this key residue (Fig. 6b). First, hOGG1 wedges the aryl ring of Tyr 203 into the space between C⁰ and the base on its 5' side (C^{−1}), thereby unstacking the two bases and creating a sharp kink in the DNA, which greatly improves access to the Watson–Crick edge of the base from the minor-groove side. C⁰ is also unstacked from the base on its 3' side (T¹), but this does not appear to be enforced overtly. Residues Arg 154 and Arg 204 converge on C⁰

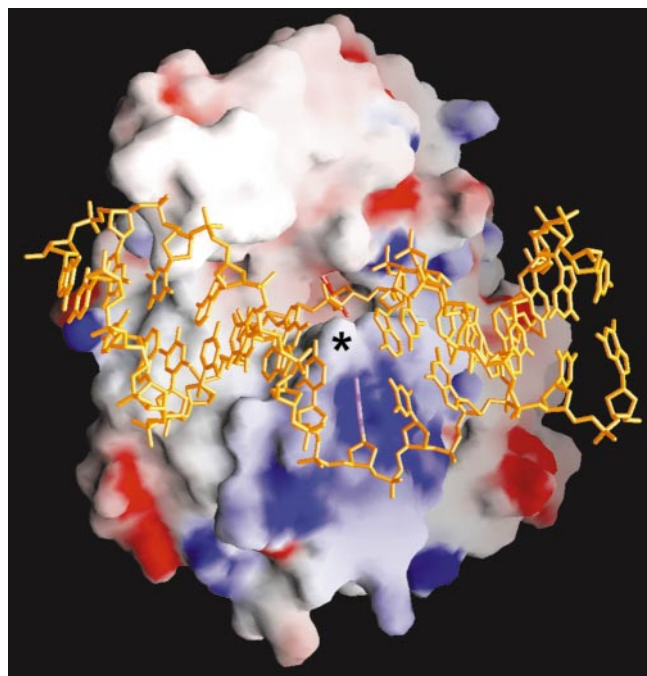


Figure 5 Electrostatic potential surface representation (GRASP⁴⁴) of hOGG1 showing the unusually neutral DNA–protein interaction surface. Negatively charged surfaces are red and positively charged surfaces blue. The oxoG base (red) is inserted into the enzyme active site, which lies underneath the top strand in this view. Asterisk denotes the side chain of Asn 149, which fills the space vacated by oxoG in the DNA helix and interacts with the estranged cytosine (magenta). The patch of blue underneath the lower strand is due to the positively charged guanidinium groups of Arg 154 and Arg 204, which contact the estranged cytosine.

from the minor-groove side, one arginine above and the other below the plane of the ring, and simultaneously establish bifurcated, bidentate hydrogen-bonding interactions with the acceptor N3 and O² atoms of the estranged base. These interactions are likely to be both exceptionally strong and specific for the presence of adjacent acceptor atoms, a feature unique to cytosine among the four DNA bases. The interaction between Asn 149 and the amide carbonyl of C⁰ completes the pentad of hydrogen bonds between the enzyme and the estranged base.

Catalytic mechanism of hOGG1

The seemingly disparate glycosylase and β -lyase activities of hOGG1 and its orthologues can be unified through a proposed mechanism^{19,31} in which Lys 249 (ref. 25) acts both to displace the oxoguanine base and to promote conjugate elimination of the 3'-phosphodiester through Schiff base chemistry (Fig. 1b). Asp 268, an invariant feature of the HhH-GPD motif, has been proposed to assist the process by transferring protons to and from Lys 249 (refs 19, 25, 27). To gain insight into the glycosylase step of the mechanism, introduced the side chain of Lys 249 computationally and swivelled the side chain of Asp 268. The salient features of this model (see <http://glviris.harvard.edu/research/hOGG1>) are that Lys 249 lies within reach (~ 2.5 Å) of the oxoG-C-1' and is near the position required for in-line back-sided displacement of the base; and that Asp 268 is suitably disposed to deprotonate and protonate Lys 249, but not while still forming a hydrogen bond with

His 270. Notably, there is no acidic residue near the oxoG base, which all but rules out the possibility that displacement of the base is accelerated by protonation.

The aminor intermediate 2 (Fig. 1b) formed by initial attack of Lys 249 rearranges to a Schiff base 3 which can be intercepted by hydride ion to yield an irreversibly linked OGG1-DNA adduct^{19,25}. This rearrangement requires deprotonation of Lys 249, presumably by Asp 268, and protonation of O-1', for which His 270 is positioned reasonably well. Such a role for His 270 could explain why this residue is invariant only in members of the HhH-GPD superfamily that catalyse the formation of Schiff base intermediates. It is also conceivable that His 270, by virtue of its exposure to solvent, may help to recharge Asp 268 for successive rounds of proton abstraction from the substrate.

In contrast to the intrinsically linked glycosylase and β -lyase activities of hOGG1 and endo III, a second class of HhH-GPD proteins are able to excise damaged bases but lack β -lyase activity; these so-called monofunctional glycosylases, represented by AlkA and MutY, simply replace the C-1' base lesion with an OH group derived from water. AlkA and other HhH-GPD monofunctional glycosylases have a residue other than lysine at the position corresponding to Lys 249, but do contain an aspartate residue corresponding to Asp 268 (ref. 25) (Fig. 3a), suggesting that both classes use the invariant aspartate residue to activate a catalytic nucleophile, which is either a water-derived hydroxide ion (monofunctional glycosylases) or the ϵ -amino group of a lysine residue (glycosylase/ β -lyases)²⁷. The structure of the hOGG1-DNA complex supports this hypothesis: the catalytic aspartate residues of hOGG1 and AlkA, Asp268 and Asp 238, respectively, are located at nearly superimposable positions in the co-crystal structure of hOGG1 and the unbound structure of AlkA²⁷. Several residues that make backbone contacts to DNA in the hOGG1 structure occupy similar positions in the free AlkA structure, suggesting that the DNA adopts a similar overall orientation in the two. Intriguingly, the two side chains between which the oxoG is bound in our structure (Fig. 6a), Cys 253 and Phe 319, correspond to Tyr 222 and Tyr 273 in AlkA, lending credence to the proposal that AlkA recognizes its electron-deficient lesions by sandwiching them between two electron-donating aromatic side-chains²⁷. Adenine soaked into crystals of *E. coli* MutY has been shown to bind in a

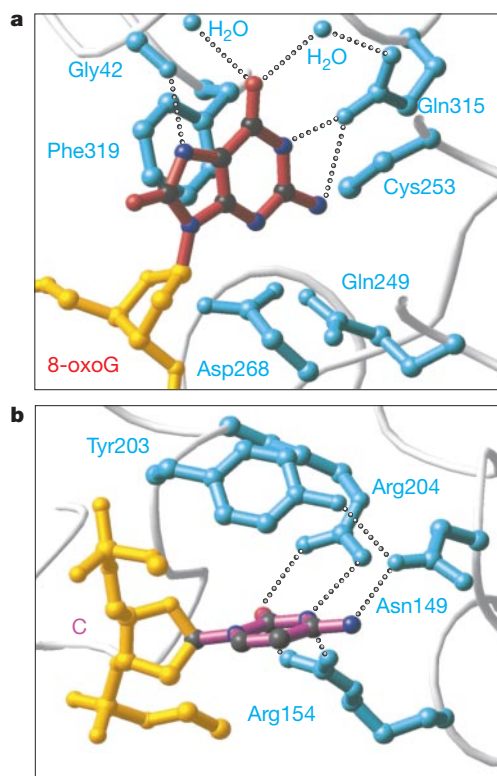


Figure 6 Recognition of the oxoG lesion and the estranged cytosine. **a**, Ball-and-stick representation of the enzyme active site, showing the residues that interact with oxoG. The oxoG base is stacked between Phe 319 and Cys 253. Residues Gly 42, Gln 315 and two water molecules hydrogen bond to the Watson-Crick and Hoogsteen faces of the lesion base. **b**, The recognition pocket for the estranged cytosine. The cytosine paired opposite oxoG is recognized by tandem bidentate hydrogen-bonding interactions with Arg 154 and Arg 204, an edge-face interaction with Tyr 203 and an additional hydrogen bond with Asn 149. The Arg-154-C⁰ and Arg-204-C⁰ contacts have nearly ideal geometry (bond angles 140–150°) and distance parameters (O²-N₁, 2.9, 3.2 Å; N3-N₁, 3.0, 3.0 Å, respectively) for bifurcated hydrogen bonds.

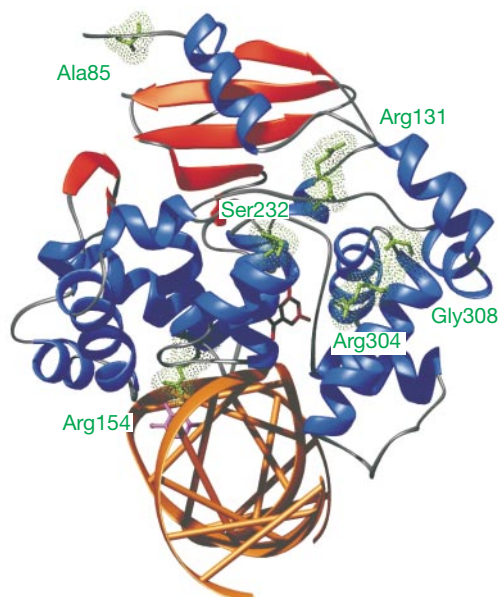


Figure 7 Ribbon diagram of the hOGG1-DNA complex with known mutant and polymorphic residues shown as green stippled surfaces.

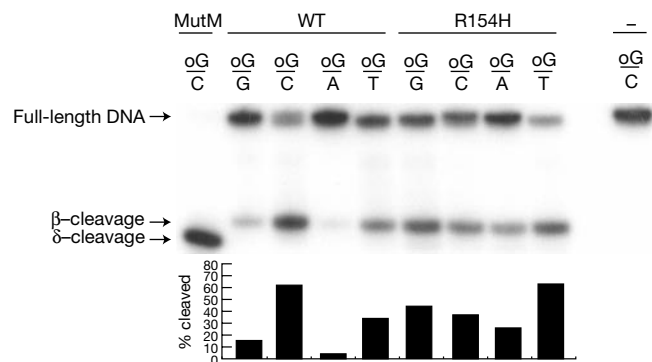


Figure 8 Enzymatic activity and substrate specificity of R154H mutant hOGG1. Purified full-length hOGG1 (WT), R154H mutant hOGG1 (R154H), the *E. coli* 8-oxoguanine glycosylase (MutM) or no enzyme (–) were incubated with a 32 P 5'-end-labelled 25-base oligonucleotide containing a centrally located oxoG (oG) residue paired opposite G, C, A or T in the complementary strand. Whereas hOGG1 yields only the characteristic β -cleavage product, the bacterial MutM produces the δ -cleavage product under these conditions; this difference allows us to rule out contamination by MutM in the recombinant OGG1 proteins. The extent of cleavage is represented by histograms below the gels.

similar manner²⁹, with the base stacked against the residue of MutY (Met 185) that corresponds to Phe 319 of hOGG1. Also, like Gln 315 of hOGG1, the sequence-related residue in MutY (Gln 182, Fig. 3) contacts the Watson–Crick face of the cognate base. This analysis suggests that there is a high probability that all HhH–GPD base-excision repair proteins bind their substrates in a similar way, using residues located at structurally related positions to recognize and repair damaged bases.

Mutations and polymorphisms in OGG1

Mutations in the *OGG1* gene have been reported in a small but significant number of transformed cells and primary tumours. The structure of the hOGG1–DNA complex provides a basis for understanding the known effects of biochemically characterized mutations and even perhaps for predicting the effects of new ones (Fig. 7). Among the mutations reported so far, an especially intriguing one was discovered through sequencing of the *HOGG1* locus in gastric cancer cell lines²³. The cell line MKN4, derived from a p53⁺

adenosquamous carcinoma, carries a mutation in the hOGG1-coding sequence that results in the conversion of Arg 154 to histidine (R154H hOGG1; Fig. 8). Arg 154 is one of the residues that check the identity of the base opposite oxoG by making direct and highly base-specific contacts to the estranged cytosine (Fig. 6b), hence any change at this position is particularly dangerous. To determine the effect of the R154H mutation on hOGG1 activity, we expressed the recombinant protein in *E. coli* and analysed its catalytic activity using DNA cleavage assays. The base-excision activity of R154H hOGG1 on oxoG–C was similar to that of the wild-type enzyme; however, R154H hOGG1 exhibited a substantial increase in its activity on oxoG opposite bases other than cytosine (Fig. 7). Especially significant from a biological standpoint is the fact that the mutant protein carries out promutagenic repair of oxoG–A nearly as efficiently as antimutagenic repair of oxoG–C.

Homozygous mutation of Arg 131 to glutamine (R131Q hOGG1) was discovered in a primary cancer of the lung, and the recombinant R131Q hOGG1 protein was found to be devoid of biochemical activity²¹. The R131Q mutation probably has an adverse effect on the folding of the protein, as the guanidine head group of Arg 131 lies on the interior of hOGG1 and participates in an extensive network of hydrogen-bonding interactions to neighbouring residues, and few if any of these are likely to form with the shorter glutamine side chain.

A Cys $\rightarrow$ Ser polymorphism at position 326 has been reported to diminish slightly the activity of the *hOGG1* gene in complementation assays in *E. coli*³², and the activity of the recombinant hOGG1 protein *in vitro*²², but no statistically significant association has been found between either allele and gastric cancer²³. Residue 326 lies within a short disordered peptide segment at the C terminus of our construct, hence its functional role is unclear. A number of other reported mutants in hOGG1 map to the surface of the protein, where they are unlikely to affect the activity of the protein substantially, and in fact may represent minor polymorphic forms rather than sporadic mutations. These include A85S (ref. 21), S232T (ref. 21) and G308E (ref. 24).

The direct involvement of hOGG1 protein in senescence was suggested by the discovery that three independent strains of mice that are genetically prone to premature ageing, somatic mutations and oxidative damage possess an inherited homozygous mutation of Arg 304 to tryptophan. Recombinant murine R304W OGG1 shows reduced base-excision activity and is highly thermolabile³³. The hydrocarbon side chain of Arg 304 is packed into the protein

Table 1 Data collection, structure determination and refinement statistics

Crystal	Native	SeMet1	SeMet2	SeMet3
Source	CHESS A1	NLSL X12C	NLSL X12C	NLSL X12C
Wavelength	0.9350	0.9790	0.9787	0.9537
Resolution (Å)	25–2.1 (2.18–2.10)	25–2.9 (3.05–2.90)	25–2.9 (3.05–2.90)	25–2.9 (3.05–2.90)
R_{sym} *	0.054 (0.384)	0.064 (0.179)	0.066 (0.173)	0.070 (0.205)
Total no. of observations	169,021	114,027	112,470	118,283
No. of unique observations	59,930 (5972)	12,608 (1023)	12,612 (1030)	12,859 (1155)
Completeness	98.8% (98.4%)	97.7% (81.1%)	97.7 (81.6%)	99.2% (92.8%)
$\langle I \rangle / \langle \sigma(I) \rangle$	16.3 (2.1)	12.5 (7.9)	11.5 (6.1)	11.8 (6.7)
No. of sites		3	3	3
Phasing power (acentric/centric)†		1.40/1.70	1.28/1.56	0.82/0.99
R_{cullis} and ‡		0.76	0.79	0.92
Resolution (Å)	25–2.1			
No. of non-hydrogen atoms	3,184			
No. of water molecules	133			
$R_{\text{working}}^{\S}$ (%)	0.2315			
$R_{\text{free}}^{\S}$ (%)	0.2629			
Average B factor	36.94			
r.m.s. deviation bonds (Å)	0.0059			
r.m.s. deviation angles (°)	1.218			
Ramachandran distribution¶	89.7, 9.9			
(% core, allowed, generous, disallowed)	0.4, 0.0			

* R_{sym} is the unweighted R value on I of symmetry-related reflections ($\sum_{hkl} |I_{hkl} - \langle I_{hkl} \rangle| / \sum_{hkl} I_{hkl}$).

† Phasing power is the r.m.s. of the heavy atom structure factor amplitudes divided by the r.m.s. of the residual lack of closure error.

‡ R_{cullis} is the mean lack of closure error divided by the isomorphous difference.

$R_{\text{working}}^{\S} = \sum_{hkl} |F_{\text{obs}} - F_{\text{calc}}| / \sum_{hkl} |F_{\text{obs}}|$; $R_{\text{free}}^{\S}$ is calculated from 10% of reflections not used in model refinement.

¶ Determined using PROCHECK⁴⁵.

core, with the polar head group extruded into a water-filled channel. The increase in steric bulk that results from the introduction of a tryptophan side chain probably disrupts optimal packing of the protein core.

Discussion

hOGG1 is a key antioxidant protein, which counters the cellular burden imposed by ROS and free radicals. The structure of hOGG1 bound to an oxoG-C substrate reveals an extrahelical active site that is tailored to the function of recognizing oxidatively damaged guanines and excluding normal DNA bases. Unexpectedly, the protein does not interact at all with the most distinctive feature of oxoG, its 8-oxo-carbonyl function, but instead detects a more subtle but no less distinctive feature, a proton on N7 capable of donating a hydrogen bond. The absence of specific recognition of the oxo-8-carbonyl contact is fortuitous, because it allows hOGG1 to recognize and repair a hydrolytically ring-opened form of guanine known as the FaPy lesion⁹. The analogous carbonyl in FaPy is displaced from the corresponding position in oxoG, whereas the analogous NHs of the two lesions are in nearly identical positions.

Compromised OGG1 function in mice is associated with premature senescence³³, an increase in levels of genomic oxoG and acceleration of the basal mutation rate¹⁶. It remains to be seen whether loss of OGG1 alone causes increased tumorigenesis in the absence of oxidative stress. However, under conditions in which markedly increased amounts of oxoG are produced, such as in lung tissue that has been chronically exposed to cigarette smoke, even wild-type levels of enzyme or the partially diminished levels brought on by chromosome 3p deletion may be insufficient to counter oxoG mutagenesis. Much attention has been focused on loss-of-function mutations in hOGG1, and the biochemical effects of these can now be understood on the basis of structure. We have reported that a gain-of-function mutation in hOGG1 might actually drive oxidative mutagenesis. Specifically, the R154H mutation affects a residue responsible for checking the cytosine opposite oxoG; this mutation alters the activity of the protein so that it repairs oxoG less efficiently than wild type and operates on oxoG-A nearly as well as oxoG-C. Replacement of wild-type hOGG1 with R154H hOGG1 could drive forward the mutational process by allowing oxoG-C to escape repair and be mis-replicated to oxoG-A, and then accelerating the conversion of oxoG-A to T-A.

Catalysing the ejection of a lesion base from duplex DNA is no straightforward task. The reaction trajectory for nucleophilic attack on C-1' of the lesion is blocked in a B-form duplex. Furthermore, both the attacking nucleophile and its target sugar must be stripped of interfering solvent water molecules, which are present in enormous molar excess. hOGG1 solves this mechanistic dilemma by extruding the entire oxoG nucleoside from the helix and engulfing it in a concave active-site pocket, thus stripping solvent from Lys 249 and the sugar of oxoG and bringing these two reacting partners together. The notion that fundamental reaction chemistry dictates the extrahelical repair strategy is supported by the fact that three families of DNA glycosylases that act on single-base lesions, HhH-GPD enzymes, uracil DNA glycosylases^{34,35} and 3-alkyadenine DNA glycosylases³⁶, all use extrahelical repair despite having completely different structures.

Having concluded that most if not all DNA glycosylases acting on single-base lesions use an extrahelical repair mechanism, it becomes especially relevant to understand how these proteins scan DNA to identify and expel their cognate lesions. Notably, the three classes of structurally distinct single-base DNA glycosylases have another feature in common: they all contact the backbone of only the lesion-containing DNA strand^{34–36}. This stands in stark contrast to the most known DNA-binding proteins, including processing enzymes such as DNA methyltransferases and restriction endonucleases, which invariably contact the backbone of both DNA strands. The unusual degree of strand bias shown by DNA glyco-

syases may hint at a new scanning process used by these proteins to locate damaged bases in the genome³⁷. □

Methods

Crystallization and data collection

A fragment of the human 8-oxoguanine DNA glycosylase (hOGG1) comprising amino acids 12–327 was cloned into the pET30a vector (Novagen) and overexpressed in *E. coli* BL21(DE3) lacking the MutM gene (H.M. Nash and G.L.V., unpublished data). After immobilizing by metal-affinity chromatography (IMAC) using Talon resin (Clontech), the N-terminal six-histidine tag was removed by enterokinase (New England Biolabs) digestion, and the protein was purified using Mono-S cation exchange and Superdex-200 gel filtration chromatography (Amersham Pharmacia Biotech). Selenomethionine-derived protein was expressed as described³⁸ and purified under stringently reducing conditions. Synthetic oligonucleotides (5'-AGCGTCCA oxoG GTCTACC and 5'-TGGTAGACCTGGACGC) were purified using denaturing PAGE. Protein (560 µM) and duplex DNA were mixed in a 1:1.5 (protein:DNA) ratio in 20 mM Tris-HCl pH 7.4, 100 mM NaCl, 1 mM EDTA and 0–10 mM dithiothreitol (DTT) at 4 °C. Hexagonal rod-shaped crystals grew in 1–3 weeks at 4 °C in hanging drops using 1 µl of complex solution mixed with 1 µl of reservoir solution: 100 mM sodium cacodylate (pH 6.5), 200 mM calcium acetate and 17–18% PEG 8000. Crystals were briefly soaked in the reservoir solution supplemented with 25% glycerol before freezing in liquid nitrogen. A high-resolution native data set was collected on the A1 beamline at CHESS using a ADSC Quantum-4 detector; MAD data were collected at the X12C beamline at NSLS using a Bragg 1k CCD detector. The crystals belong to space group $P6_322$ ($a = b = 92.4$ Å, $c = 212.9$ Å).

Phasing and refinement

All X-ray data were indexed and scaled using DENZO/SCALEPACK³⁹. The positions of three selenium atoms from the MAD data were located and the initial electron-density map calculated using the program SOLVE⁴⁰. Heavy-atom parameter refinement and density modification were performed using the CNS package⁴¹. An initial model was built into the electron-density map using QUANTA98 (Molecular Simulations) and refined using CNS. Rounds of model refinement were evaluated by monitoring the free R factor⁴² and stereochemical parameters. The final model contains residues 12–325 and includes three ordered N-terminal residues originating from the expression vector (Gly 9, Ser 10, Glu 11). Amino acids in an unordered loop (80–82) were excluded from the model. Amino acids Glu 71, Gln 84, Arg 87, Gln 125, Gln 128, Gln 164, Gln 227, Lys 289 and Gln 325 were modelled as alanines, because no side-chain electron density was evident in the final map. The terminal 5' residue of each oligonucleotide strand was unordered and not included in the final model.

Cleavage assays on R154H hOGG1

The R154H mutation was introduced into the expression vector encoding full-length hOGG1, and the protein was expressed in *E. coli* BL21(DE3). The protein was purified to homogeneity and normalized to the wild-type enzyme using the Bradford assay. Cleavage experiments were carried out as described¹⁹.

Received 9 December 1999; accepted 2 February 2000.

- Lindahl, T. Instability and decay of the primary structure of DNA. *Nature* **362**, 709–715 (1993).
- Lindahl, T. & Wood, R. D. Quality control by DNA repair. *Science* **286**, 1897–1905 (1999).
- Michaels, M. L. & Miller, J. H. The GO system protects organisms from the mutagenic effect of the spontaneous lesion 8-hydroxyguanine (7,8-dihydro-8-oxoguanine). *J. Bacteriol.* **174**, 6321–6325 (1992).
- Grollman, A. P. & Moriya, M. Mutagenesis by 8-oxoguanine: an enemy within. *Trends Genet.* **9**, 246–249 (1993).
- Hollstein, M. *et al.* Somatic point mutations in the p53 gene of human tumors and cell lines: updated compilation. *Nucleic Acids Res.* **24**, 141–146 (1996).
- Lu, R., Nash, H. M. & Verdine, G. L. A mammalian DNA repair enzyme that excises oxidatively damaged guanines maps to a locus frequently lost in lung cancer. *Curr. Biol.* **7**, 397–407 (1997).
- Arai, K. *et al.* Cloning of a human homolog of the yeast OGG1 gene that is involved in the repair of oxidative DNA damage. *Oncogene* **14**, 2857–2861 (1997).
- Roldan-Arjona, T. *et al.* Molecular cloning and functional expression of a human cDNA encoding the antimutator enzyme 8-hydroxyguanine-DNA glycosylase. *Proc. Natl Acad. Sci. USA* **94**, 8016–8020 (1997).
- Bjoras, M. *et al.* Opposite base-dependent reactions of a human base excision repair enzyme on DNA containing 7,8-dihydro-8-oxoguanine and abasic sites. *EMBO J.* **16**, 6314–6322 (1997).
- Rosenquist, T. A., Zharkov, D. O. & Grollman, A. P. Cloning and characterization of a mammalian 8-oxoguanine DNA glycosylase. *Proc. Natl Acad. Sci. USA* **94**, 7429–7434 (1997).
- Radicella, J. P., Dherin, C., Desmaziere, C., Fox, M. S. & Boiteux, S. Cloning and characterization of hOGG1, a human homolog of the OGG1 gene of *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA* **94**, 8010–8015 (1997).
- Kuo, F. C. & Sklar, J. Augmented expression of a human gene for 8-oxoguanine DNA glycosylase (MutM) in B lymphocytes of the dark zone in lymph node germinal centers. *J. Exp. Med.* **186**, 1547–1556 (1997).
- Aburatani, H. *et al.* Cloning and characterization of mammalian 8-hydroxyguanine-specific DNA glycosylase/apurinic, apyrimidinic lyase, a functional mutM homologue. *Cancer Res* **57**, 2151–2156 (1997).
- van der Kemp, P. A., Thomas, D., Barbey, R., de Oliveira, R. & Boiteux, S. Cloning and expression in *Escherichia coli* of the OGG1 gene of *Saccharomyces cerevisiae*, which codes for a DNA glycosylase that excises 7,8-dihydro-8-oxoguanine and 2,6-diamino-4-hydroxy-5-N-methylformamidopyrimidine. *Proc. Natl Acad. Sci. USA* **93**, 5197–5202 (1996).

15. Bruner, S. D., Nash, H. M., Lane, W. S. & Verdine, G. L. Repair of oxidatively damaged guanine in *Saccharomyces cerevisiae* by an alternative pathway. *Curr. Biol.* **8**, 393–403 (1998).
16. Klungland, A. *et al.* Accumulation of premutagenic DNA lesions in mice defective in removal of oxidative base damage. *Proc. Natl Acad. Sci. USA* **96**, 13300–13305 (1999).
17. Takao, M., Aburatani, H., Kobayashi, K. & Yasui, A. Mitochondrial targeting of human DNA glycosylases for repair of oxidative DNA damage. *Nucleic Acids Res.* **26**, 2917–2922 (1998).
18. Nishioka, K. *et al.* Expression and differential intracellular localization of two major forms of human 8-oxoguanine DNA glycosylase encoded by alternatively spliced OGG1 mRNAs. *Mol. Biol. Cell* **10**, 1637–1652 (1999).
19. Nash, H. M. *et al.* Cloning of a yeast 8-oxoguanine DNA glycosylase reveals the existence of a base-excision DNA-repair protein superfamily. *Curr. Biol.* **6**, 968–980 (1996).
20. Shinmura, K., Kasai, H., Sasaki, A., Sugimura, H. & Yokota, J. 8-Hydroxyguanine (7,8-dihydro-8-oxoguanine) DNA glycosylase and AP lyase activities of hOGG1 protein and their substrate specificity. *Mutat. Res.* **385**, 75–82 (1997).
21. Chevillard, S. *et al.* Mutations in OGG1, a gene involved in the repair of oxidative DNA damage, are found in human lung and kidney tumours. *Oncogene* **16**, 3083–3086 (1998).
22. Dherin, C., Radicella, J. P., Dizdaroğlu, M. & Boiteux, S. Excision of oxidatively damaged DNA bases by the human alpha-hOgg1 protein and the polymorphic alpha-hOgg1(Ser326Cys) protein which is frequently found in human populations. *Nucleic Acids Res.* **27**, 4001–4007 (1999).
23. Shinmura, K. *et al.* Infrequent mutations of the hOGG1 gene, that is involved in the excision of 8-hydroxyguanine in damaged DNA, in human gastric cancer. *Jpn. J. Cancer Res.* **89**, 825–828 (1998).
24. Blons, H. *et al.* Frequent allelic loss at chromosome 3p distinct from genetic alterations of the 8-oxoguanine DNA glycosylase 1 gene in head and neck cancer. *Mol. Carcinog.* **26**, 254–260 (1999).
25. Nash, H. M., Lu, R., Lane, W. S. & Verdine, G. L. The critical active-site amine of the human 8-oxoguanine DNA glycosylase, hOgg1: direct identification, ablation and chemical reconstitution. *Chem. Biol.* **4**, 693–702 (1997).
26. Thayer, M. M., Ahern, H., Xing, D., Cunningham, R. P. & Tainer, J. A. Novel DNA binding motifs in the DNA repair enzyme endonuclease III crystal structure. *EMBO J.* **14**, 4108–4120 (1995).
27. Labahn, J. *et al.* Structural basis for the excision repair of alkylation-damaged DNA. *Cell* **86**, 321–329 (1996).
28. Kuo, C. F. *et al.* Atomic structure of the DNA repair [4Fe-4S] enzyme endonuclease III. *Science* **258**, 434–440 (1992).
29. Guan, Y. *et al.* MutY catalytic core, mutant and bound adenine structures define specificity for DNA repair enzyme superfamily. *Nature Struct. Biol.* **5**, 1058–1064 (1998).
30. Yamagata, Y. *et al.* Three-dimensional structure of a DNA repair enzyme, 3-methyladenine DNA glycosylase II, from *Escherichia coli*. *Cell* **86**, 311–319 (1996).
31. Dodson, M. L., Michaels, M. L. & Lloyd, R. S. Unified catalytic mechanism for DNA glycosylases. *J. Biol. Chem.* **269**, 32709–32712 (1994).
32. Kohno, T. *et al.* Genetic polymorphisms and alternative splicing of the hOGG1 gene, that is involved in the repair of 8-hydroxyguanine in damaged DNA. *Oncogene* **16**, 3219–3225 (1998).
33. Choi, J. Y. *et al.* Thermolabile 8-hydroxyguanine DNA glycosylase with low activity in senescence-accelerated mice due to a single-base mutation. *Free Radical Biol. Med.* **27**, 848–854 (1999).
34. Barrett, T. E. *et al.* Crystal structure of a thwarted mismatch glycosylase DNA repair complex. *EMBO J.* **18**, 6599–6609 (1999).
35. Parikh, S. S. *et al.* Base excision repair initiation revealed by crystal structures and binding kinetics of human uracil-DNA glycosylase with DNA. *EMBO J.* **17**, 5214–5226 (1998).
36. Lau, A. Y., Schärer, O. D., Samson, L., Verdine, G. L. & Ellenberger, T. Crystal structure of a human alkylbase-DNA repair enzyme complexed to DNA: mechanisms for nucleotide flipping and base excision. *Cell* **95**, 249–258 (1998).
37. Verdine, G. L. & Bruner, S. D. How do DNA repair proteins locate damaged bases in the genome? *Chem. Biol.* **4**, 329–334 (1997).
38. Van Duyne, G. D., Standaert, R. F., Karplus, P. A., Schreiber, S. L. & Clardy, J. Atomic structures of the human immunophilin FKBP-12 complexes with FK506 and rapamycin. *J. Mol. Biol.* **229**, 105–124 (1993).
39. Otwinowski, Z. M. W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1996).
40. Terwilliger, T. C. & Berendzen, J. Automated MAD and MIR structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **55**, 849–61 (1999).
41. Brunger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
42. Brunger, A. Assessment of phase accuracy by cross validation: the free R value. Methods and applications. *Acta Crystallogr.* **49**, 24–36 (1993).
43. Carson, M. J. Ribbons 2.0. *J. Appl. Crystallogr.* **24**, 379–384 (1991).
44. Nicholls, A., Sharp, K. A., Honig, B. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins Struct. Funct. Genet.* **11**, 281–296 (1991).
45. Laskowski, R. J., MacArthur, M. W., Moss, D. S. & Thornton, J. M. PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **26**, 283–290 (1993).

Acknowledgements

We thank E. Lawson for assisting in DNA synthesis and protein expression; H. Huang, Q. Fan, L. Chen, S. Ealick, A. Lau, S. Liemann, H. Nash and members of the Verdine, S. C. Harrison and D.C. Wiley laboratories for helpful discussions; C. Schafmeister, R. Crouse, M. Przetakiewicz, R. Sweet, C. Metz and the staff of the CHESS A1 and NSLS X12C synchrotron beamlines for invaluable help with data collection; and K. Haushalter and X. Zhang for critical reading of this manuscript. This research was funded by the NIH.

Correspondence and requests for materials should be addressed to G.L.V. (e-mail: verdine@chemistry.harvard.edu). Atomic coordinates have been deposited in the PDB under access code 1EBM and can be downloaded from <http://glivis.harvard.edu/research/HOGG1>.

Formation of molecular gas in the tidal debris of violent galaxy–galaxy interactions

Jonathan Braine*, Ute Lisenfeld†, Pierre-Alain Duc‡ & Stéphane Leon§

* Observatoire de Bordeaux, UMR 5804, CNRS/INSU, BP89, F-33270 Floirac, France

† Institut de Radioastronomie Millimétrique, Avenida Divina Pastora 7, NC18012 Granada, Spain

‡ Institute of Astronomy, Madingley Road, Cambridge CB3 0HA, UK and CNRS and CEA/DSM/DAPNIA Service d'astrophysique, Saclay, 91191 Gif sur Yvette cedex, France

§ ASIAA, Academia Sinica, PO Box 1-87, Nanking, Taipei 115, Taiwan

In many gravitational interactions between galaxies, gas and stars that have been torn from the precursor galaxies can collect in tidal 'tails'. Star formation begins anew in some of these regions, producing tidal dwarf galaxies^{1–4}. Observations of these new galaxies provides insight into processes relevant to galaxy formation more generally, because the timescale of the interaction is well defined. But tracking the star formation process has hitherto been difficult because the tidal dwarf galaxies with young stars showed no evidence of the molecular gas out of which those young stars formed^{5–8}. Here we report the discovery of molecular hydrogen (traced by carbon monoxide emission) in two tidal

dwarf galaxies. In both cases, the concentration of molecular gas peaks at the same location as the maximum in atomic-hydrogen density, unlike the situation in most gas-rich galaxies. We infer from this that the molecular gas formed from the atomic hydrogen, rather than being torn in molecular form from the interacting galaxies. Star formation in the tidal dwarf galaxies therefore appears to mimic the process in normal spiral galaxies like our own.

Tidal dwarf galaxies (TDGs) are gas-rich irregular galaxies made out of stellar and gaseous material pulled out by tidal forces from the disks of the colliding parent galaxies into the intergalactic medium^{9–12}. They are found at the ends of long tidal tails, sometimes 100 kpc from the nuclei of their progenitors, and host active star-forming regions. TDGs contain two main stellar components: young stars recently formed by collapse of expelled atomic hydrogen (H I) clouds, and an older stellar population, at least 1 Gyr old, originally part of the disk of the parent galaxies. Their overall gaseous and stellar properties range between those of classical dwarf irregular galaxies and blue compact dwarf galaxies, with the exception of their metallicity which is higher—typical of the outer disk of a spiral galaxy¹². Whether a large fraction of dwarf galaxies were formed through tidal encounters in the early Universe—when spiral galaxies were more gaseous and less metal-rich, and collisions more frequent—is an open question, and one of the reasons to study TDGs. One way to answer this question is the dark-matter content of dwarf galaxies. Observations of ordinary dwarf galaxies show that a lot of dark matter, or mass that is in some as yet invisible form, is necessary to account for their rotation velocities. Numerical simulations of gravitational interactions indicate that TDGs should have very little dark matter¹³ if the dark matter is, as currently believed, in the form of a large halo and not in, say, a

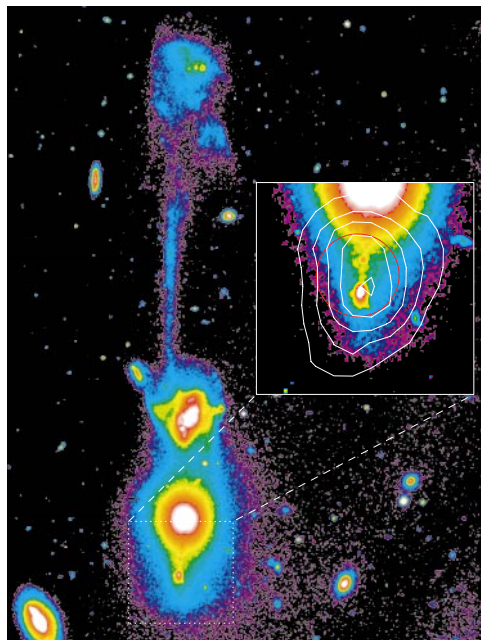


Figure 1 The southern Tidal Dwarf Galaxy (shown in the magnified view) in the interacting system Arp105. (The latter is also known as NGC3561 (ref. 24) and the "Guitar"). H I emission contours are superposed on a magnified view of a V-band image of A105S (ref. 3). The frame is $4.4' \times 5.9'$; north is up and east to the left. Red circle is (0,0) position of CO observations, and represents the full-width at half-maximum FWHM ($22''$) of the CO(1–0) beam. The Arp105 system^{2,3} is an interaction between a spiral and an elliptical which has generated an H I-rich extended TDG at the end of the northern tidal tail and a more compact TDG at the tip of the southern tail from the spiral. A105S contains roughly $5 \times 10^8 M_{\odot}$ of H I and strong H α emission, corresponding to a star-formation rate of $\sim 0.2 M_{\odot} \text{ yr}^{-1}$. Nonetheless, stellar synthesis models²⁵ of A105S indicate that the stellar mass is dominated by the old spiral disk population while the luminosity comes mostly from stars formed *in situ*.

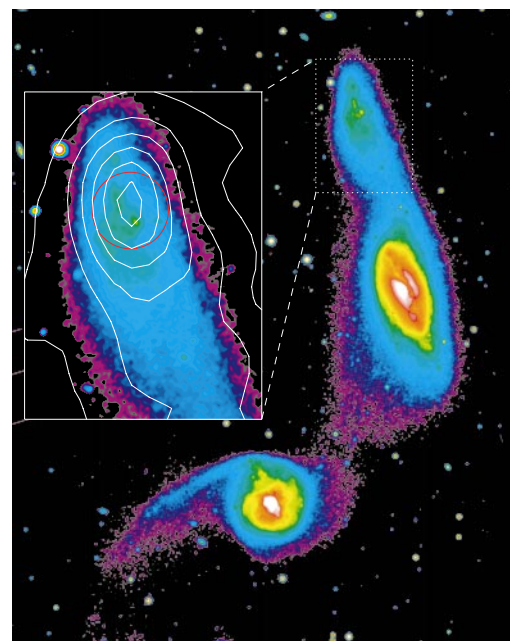


Figure 2 The tidal dwarf galaxy (shown in the magnified view) in the interacting system Arp245 (NGC2992/3²⁴). H I emission contours are superposed on magnified view of a V-band image of A245N (ref. 14). The frame is $5.8' \times 7.4'$; north is up and east to the left. Red circle is (0,0) position of CO observations and represents the FWHM ($22''$) of the CO(1–0) beam. Arp245 is an interaction between two spirals. The TDG, A245N, has been formed in the tidal tail which stems from NGC2992 and contains nearly twice as much H I as A105S but slightly weaker H α emission. The old stellar population is more prominent in A245N than in A105S (ref. 14). The physical size and total H I mass of Arp245 are smaller than in Arp105.

Table 1 Properties of the Arp105 and Arp245 TDGs

	A105S	A245N
RA(J200)	11 h 11 min 13.5 s	09 h 45 min 44.1 s
Dec(J2000)	+28° 41' 20"	−14° 17' 28"
H I velocity (LSR)	$cz = 8,890 \text{ km s}^{-1}$	$cz = 2,175 \text{ km s}^{-1}$
Adopted distance	115 Mpc	31 Mpc
$L_{\text{H}\alpha}$	$(1-2) \times 10^{40} \text{ erg s}^{-1}$	$7 \times 10^{39} \text{ erg s}^{-1}$
$M_B, L_B/L_{B\odot}$	−16.9, 9×10^8	−17.25, 1.2×10^9
$B - V$	0.3	0.55
$M_{\text{H I}}$	$5 \times 10^8 M_\odot$	$9 \times 10^8 M_\odot$
M_{H_2}	$\geq 2.2 \times 10^9 M_\odot$	$\geq 1.4 \times 10^9 M_\odot$

Data from refs 2 and 3 for A105S, and ref. 14 for A245N. Position is (0,0) of CO map, and velocity is zero of spectra (Fig. 3). M_B and L_B include a correction for galactic absorption of 0.3 mag for A245N. A105S is at high Galactic latitude so no correction is applied. The molecular gas mass is estimated using a $N(\text{H}_2)/\text{CO}$ factor of $2 \times 10^{20} \text{ K km s}^{-1} \text{ cm}^{-2}$ and probably represents a lower limit because weaker, undetected, CO emission may be present at other positions. We have included the mass of helium in the molecular clouds. Relative to the velocities of the TDGs, the spiral and elliptical in the Arp105 system have velocities of ~ 130 and $\sim 400 \text{ km s}^{-1}$. In Arp245, the velocities of the spirals NGC2992 and NGC2993 are 155 and 245 km s^{-1} with respect to the TDG. LSR, local standard of rest; $L_{\text{H}\alpha}$, luminosity of the $\text{H}\alpha$ line; L_B , luminosity in blue band; $L_{B\odot}$, luminosity in blue band of the Sun; B , blue band magnitude; V , visual band magnitude; M_B , blue band absolute magnitude; $M_{\text{H I}}$, mass of H I gas; M_{H_2} , mass of H_2 gas.

rotating disk. Thus, if TDGs are found to possess the same dark-matter properties as other dwarf galaxies, then powerful constraints are placed on the form of dark matter. If, on the other hand, TDGs do not contain such dark matter, then tidal interactions cannot be the principal formation mechanism for these small galaxies, nor can dark matter be part of galactic disks.

The observations were performed with the 30-m telescope operated by the Institut de Radioastronomie Millimétrique (IRAM) on Pico Veleta, Spain, in June 1999. Carbon monoxide (CO) emission was detected in the southern TDG in Arp105 (Fig. 1; hereafter A105S) and main TDG in Arp245 (Fig. 2; hereafter A245N) in both the ground state $\text{CO}(J=1-0)$ and the $\text{CO}(J=2-1)$ transitions. Small maps were made of both sources to localize the CO emission with respect to the atomic hydrogen (H I), ionized gas ($\text{H}\alpha$), and optical continuum^{3,14}. The central (0,0) $\text{CO}(1-0)$ spectra are shown in Fig. 3 along with the H I spectra at the same positions with a similar beamsize. The $\text{CO}(1-0)$ luminosities and derived H_2 mass estimates (see Table 1) of A105S and A245N are far greater than those of other dwarf galaxies¹⁵. Despite the different environments, the star-formation efficiency, defined as the rate of star formation per mass of molecular gas, is quite close to that observed in the Milky Way and other spiral galaxies¹⁶.

Small CO maps have been made consisting of six positions towards A105S and four positions towards A245N (see Table 2). In both cases, the CO peaks at the H I column density maximum and

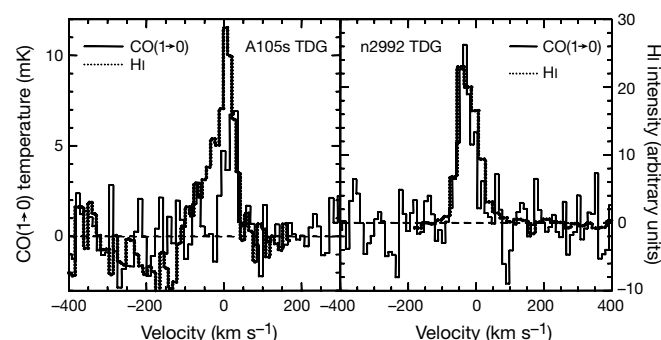


Figure 3 $\text{CO}(1-0)$ and H I spectra of the (0,0) position of tidal dwarf galaxies A105S (left) and A245N (right). The velocities and line widths of the CO and H I emission are very similar. Towards the very compact TDG A105S, the CO emission is not resolved. The CO emission in A245N is extended with detections in at least 3 of 4 observed points. The $\text{H}\alpha$ emission in A245N (ref. 14) decreases substantially towards the (0,−10) offset position while the H I (ref. 14) and CO (see Table 2) are still strong. In contrast, the $\text{H}\alpha$ emission towards the (3,14) position is comparable to the centre whereas the CO and H I have decreased significantly. H I intensity is in arbitrary units.

Table 2 Molecular gas in TDGs

Source	Offset ($\delta\text{RA}, \delta\text{dec.}$)	I_{CO} (K km s^{-1})	Noise (mK)	Vel. (km s^{-1})	ΔV_{FWHM} (km s^{-1})
A105S	(0,0)	0.3 ± 0.05	2.5	19 ± 6	38 ± 10
		0.2 ± 0.05	3.5	16 ± 5	25 ± 10
A105S	(10,0)	0.1 ± 0.5	3	59 ± 6	19 ± 10
A105S	(−10,0)	0.15 ± 0.05	2.8	$−4 \pm 5$	22 ± 13
A105S	Other	0.15 ± 0.05	2.5	15 ± 6	15 ± 8
A245N	(0,0)	1.3 ± 0.1	4.9	$−32 \pm 4$	48 ± 10
		2.0 ± 0.5	18	$−32 \pm 10$	66 ± 20
A245N	(3,14)	0.6 ± 0.15	7	$−33 \pm 10$	47 ± 15
A245N	(0,−10)	0.9 ± 0.15	7.6	$−27 \pm 7$	44 ± 14
A245N	(0,−20)	0.5 ± 0.2	9.5	$−27 \pm 12$	32 ± 15

The offset is in arcsec with respect to the position given in Table 1 and the red circle in Figs 1 and 2. I_{CO} is the flux of the CO line expressed in K km s^{-1} . The lines without source and offset are the $\text{CO}(2-1)$ observations of the preceding source and position. Noise levels (r.m.s.) are given for channel widths of 2 MHz in the $\text{CO}(1-0)$ line and 2.5 MHz in the $\text{CO}(2-1)$ line. The velocity of the line centre is with respect to the H I velocity given in Table 1. Note that the 'detections' of the off-centre A105S positions are uncertain, so the velocities and line widths may be meaningless. The spectra for each position were averaged, a continuum level was then subtracted such that the average flux outside the line window is zero, and resulting spectra were smoothed to yield the results presented in Fig. 3 and Table 2. No baselines other than the continuum level are subtracted from the data. The A105S 'other' position represents the average of the spectra for the (0,5) (0,−5) and (0,−10) positions. Taken individually, these points were not detected and yield 3σ limits of $I_{\text{CO}} \leq 0.3 \text{ K km s}^{-1}$. The CO emission from A105S is consistent with a punctual source, much like for the optical and H I. The angular resolutions are respectively 22" and 11"; beam efficiencies are 0.72 and 0.48.

the dynamics of the atomic and molecular components are virtually identical (Fig. 3). In spiral galaxies, on the other hand, H I and CO have very different distributions (see for example, refs 17 and 18), showing that the molecular gas that we have found in the TDGs has not simply been torn off the parent galaxies together with the H I, but rather has formed *in situ*. Although the calculations were performed for post-shock gas, an estimate of the molecule formation time is $t \approx n^{-1} \text{ Gyr}$ (ref. 19) where n is the density of the atomic medium in particles per cm^3 . Numerical simulations of Arp245 (ref. 14) yield an age of about 100 Myr for A245N; a rough age estimate for A105S can be obtained by dividing the projected distance to the spiral by the relative radial velocity, yielding about 200 Myr (ref. 3), sufficient for H_2 formation in standard atomic hydrogen clouds ($\bar{n} \approx 10 \text{ cm}^{-3}$). The dust on which the H_2 forms is captured from the parent galaxies and is present in the atomic gas^{20–22}. We conclude that the molecular gas has formed inside the H I clouds, and star formation is proceeding in a standard way from the molecular gas.

Our observations show that the molecular gas is an important component in the visible mass budget of TDGs, between $\sim 20\%$ and $\geq 50\%$ of the atomic hydrogen mass (see Table 1). The fact that we detect large quantities of molecular gas, and that we have every reason to believe that this gas is the result of conversion from H I into H_2 , indicates that the central regions of these objects should be gravitationally bound. If the H I were dense enough pre-encounter, then CO would form and be routinely detected beyond the optical radius in galactic disks, like H I—but it is not^{17,20,21,23}. While it was clear that TDGs are kinematically decoupled from their parent galaxies, the evidence that TDGs were bound was (until now) morphological—the accumulation of matter at the tips of the tidal tails, and the presence of star-forming regions. Although measurements with higher angular resolution are necessary, our conclusion—that TDGs are bound—provides firmer ground for the calculation of the dynamical mass; this calculation relies on the assumption that the object is gravitationally bound and in equilibrium. A more certain dynamical mass will lead to a more certain determination of dark matter—which (by definition) is detected by a discrepancy between the velocities expected on the basis of the mass of what we see directly and those velocities in fact observed. $\square$

Received 10 August; accepted 8 December 1999.

- Mirabel, I. F., Dottori, H. & Lutz, D. Genesis of a dwarf galaxy from the debris of the Antennae. *Astron. Astrophys.* **256**, L19–L23 (1992).
- Duc, P.-A. & Mirabel, I. F. Recycled galaxies in the colliding system Arp 105. *Astron. Astrophys.* **289**, 83–93 (1994).

3. Duc, P.-A., Brinks, E., Wink, J. E. & Mirabel, I. F. Gas segregation in the interacting system Arp 105. *Astron. Astrophys.* **326**, 537–553 (1997).
4. Duc, P.-A. & Mirabel, I. G. Young tidal dwarf galaxies around the gas-rich disturbed lenticular NGC 5291. *Astron. Astrophys.* **333**, 813–826 (1998).
5. Brouillet, N., Henkel, C. & Baudry, A. Detection of an intergalactic molecular complex? *Astron. Astrophys.* **262**, L5–L8 (1992).
6. Walter, F. & Heithausen, A. The discovery of a molecular complex in the tidal arms near NGC 3077. *Astron. J.* **519**, L69–L72 (1999).
7. Smith, B. J. & Higdon, J. L. A search for CO(1–0) emission from the tidal structures of interacting and merging galaxies. *Astron. J.* **108**, 837–843 (1994).
8. Smith, B. J., Struck, C., Kenney, J. D. P. & Jogee, S. The molecule-rich tail of the peculiar galaxy NGC 2782 (Arp 215). *Astron. J.* **117**, 1237–1248 (1999).
9. Zwicky, F. Multiple galaxies. *Ergebnisse Exakten Naturwissenschaften* **29**, 344–385 (1956).
10. Schweizer, F. in *Structure and Properties of Nearby Galaxies* (eds Berkhuysen, E. M. & Wielebinski, R.) 279–284 (Riedel, Dordrecht, 1978).
11. Hibbard, J. E. & van Gorkom, J. H. HI, HII, and R-Band observations of a galactic merger sequence. *Astron. J.* **111**, 655–695 (1996).
12. Duc, P.-A. & Mirabel, I. F. in *Galaxy Interactions at Low and High Redshift* (eds Barnes, J. & Sanders, D.) 61–70 (IAU Symp. 186, Kluwer, Dordrecht, 1997).
13. Barnes, J. E. & Hernquist, L. Formation of dwarf galaxies in tidal tails. *Nature* **360**, 715–717 (1992).
14. Duc, P.-A. *et al.* The interacting system NGC 2992/3 (Arp 245). *Astron. J.* (submitted).
15. Taylor, C. L., Kobulnicky, H. A. & Skillman, E. D. CO emission in low-luminosity, HI-rich galaxies. *Astron. J.* **116**, 2746–2756 (1998).
16. Kennicutt, R. C. Jr Star formation in galaxies along the Hubble sequence. *Annu. Rev. Astron. Astrophys.* **36**, 189–232 (1998).
17. Guélin, M. *et al.* 1.3 mm emission in the disk of NGC 891: Evidence of cold dust. *Astron. Astrophys.* **279**, L37–L40 (1993).
18. Braine, J., Combes, F. & Van Driel, W. NGC 4414: A flocculent galaxy with a high gas surface density. *Astron. Astrophys.* **280**, 451–467 (1993).
19. Hollenbach, D. & McKee, C. F. Molecule formation and infrared emission in fast interstellar shocks. III—results for J shocks in molecular clouds. *Astrophys. J.* **342**, 306–336 (1989).
20. Neining, N., Guélin, M., Garcia-Burillo, S., Zylka, R. & Wielebinski, R. Cold dust and molecular line emission in NGC 4565. *Astron. Astrophys.* **310**, 725–736 (1996).
21. Dumke, M. *et al.* the interstellar medium in the edge-on galaxy NGC 5907. Cold dust and molecular line emission. *Astron. Astrophys.* **325**, 124–134 (1997).
22. Braine, J. *et al.* Gas and dust in the active spiral galaxy NGC 3079. *Astron. Astrophys.* **326**, 963–975 (1997).
23. Sage, L. J. The properties and origins of molecular gas in the lenticular galaxies NGC 404, 4710 and 5195. *Astron. Astrophys.* **239**, 125–136 (1990).
24. Arp, H. Atlas of peculiar galaxies. *Astrophys. J. Suppl. Ser.* **14**, 1–20 (1966).
25. Fritz-v. Alvensleben, U. & Duc, P.-A. in *The Magellanic Clouds and other Dwarf Galaxies* (eds Braun, J. M. & Richtler, T.) 141–145 (Proceedings of the Workshop of the Graduiertenkolleg Bonn-Bochum, Shaker, Aachen, 1998).

Correspondence and requests for materials should be addressed to J.B.
(e-mail: Jonathan.Braine@obs.uv-bordeaux.fr).

Geometric quantum computation using nuclear magnetic resonance

Jonathan A. Jones^{*†}, Vlatko Vedral^{*}, Artur Ekert^{*} & Giuseppe Castagnoli[‡]

^{*} Centre for Quantum Computation, Clarendon Laboratory, Parks Road, Oxford OX1 3PU, UK

[†] Oxford Centre for Molecular Sciences, New Chemistry Laboratory, South Parks Road, Oxford OX1 3QT, UK

[‡] Elsas, Via Puccini 2, 1615 Genova, Italy

A significant development in computing has been the discovery¹ that the computational power of quantum computers exceeds that of Turing machines. Central to the experimental realization of quantum information processing is the construction of fault-tolerant quantum logic gates. Their operation requires conditional quantum dynamics, in which one sub-system undergoes a coherent evolution that depends on the quantum state of another sub-system²; in particular, the evolving sub-system may acquire a conditional phase shift. Although conventionally dynamic in origin, phase shifts can also be geometric^{3,4}. Conditional geometric (or 'Berry') phases depend only on the geometry of the path executed, and are therefore resilient to certain types of errors; this suggests the possibility of an intrinsically fault-

tolerant way of performing quantum gate operations. Nuclear magnetic resonance techniques have already been used to demonstrate both simple quantum information processing^{5–9} and geometric phase shifts^{10–12}. Here we combine these ideas by performing a nuclear magnetic resonance experiment in which a conditional Berry phase is implemented, demonstrating a controlled phase shift gate.

Any quantum computation can be built out of simple operations involving only one or two quantum bits (qubits)¹³. A particularly simple two-qubit gate in many experimental implementations, such as nuclear magnetic resonance (NMR)¹⁴, is the controlled phase shift. This may be achieved using a conditional Berry phase, and thus quantum geometrical phases can form the basis of quantum computation. We will use spin half nuclei as an example to demonstrate the feasibility of this approach, but the basic idea is general. In our experiments the state of one spin determines the Berry phase acquired by the other spin.

Suppose that a spin half nucleus undergoes a conical evolution with cone angle θ . Then the Berry phase is simply $\gamma = \pm \frac{1}{2}\Omega = \pm \pi(1 - \cos\theta)$ where the $\pm$ signs depend on whether the system is in the eigenstate aligned with or against the field, and Ω is the solid angle subtended by the conical circuit. We note that any deformation of the path of the spin which preserves this solid angle leaves the phase unchanged. Thus the phase is not affected by the speed with which the path is traversed; nor is it very sensitive to random fluctuations about the path.

Berry phases can be conveniently demonstrated in an NMR experiment¹⁵ by working in a rotating frame. We consider an ensemble of spin half particles in a magnetic field, B_0 , aligned along the z-axis; their precession frequency is then given by the Larmor frequency, ω_0 . If the spins are irradiated by a circularly polarized radio-frequency field, B_1 , at a frequency ω_{rf} , the total hamiltonian (neglecting relaxation) may be written in the rotating frame as $H = (\omega_0 - \omega_{rf})I_z + \omega_1 I_x$, where, following conventional NMR practice, the hamiltonian is described in product operator notation¹⁶ and the field strengths are written in terms of their corresponding Larmor frequencies.

When $|\omega_1| \ll |\omega_0 - \omega_{rf}|$ the hamiltonian lies close to the z-axis, while when $|\omega_1| \gg |\omega_0 - \omega_{rf}|$, the hamiltonian lies close to the x-axis. If radio-frequency radiation is applied far from resonance, the system is effectively quantized along the z-axis, and if the radio frequency is swept towards resonance ($\omega_{rf} = \omega_0$), the effective hamiltonian rotates from the z-axis towards the x-axis. If the frequency sweep is applied adiabatically then the spin will follow the hamiltonian. Next, a circular motion can be imposed by adiabatically varying the phase of the radio frequency. When the hamiltonian returns to the x-axis the frequency sweep may be reversed, so that the spin returns to its original state, aligned along the z-axis. The Berry phase acquired in this cyclic process is $\pm\pi$. If the radio-frequency field is not swept all the way to resonance, but only to some final value ω_f , the hamiltonian ends at some angle to the z-axis, and so circuits with arbitrary cone angles can be implemented. A similar case occurs if the frequency sweep is replaced by an amplitude sweep, in which the radio frequency is always applied away from resonance, and its amplitude is raised smoothly from zero to some final value, ω_1 .

This situation arises naturally in a system of two weakly coupled spins, I and S . For simplicity we consider a heteronuclear system, so that ω_I and ω_S are very different, and only one spin (say I) is close to resonance. The two transitions of I (corresponding to the two possible states of spin S) will be split by $\pm\pi J$, and so will have different resonance offsets. After an amplitude sweep the orientation of the effective hamiltonian depends on the resonance offset, and so θ (and hence the Berry phase acquired) will depend on the state of spin S . This permits a conditional Berry phase to be applied to spin I , where the size of the phase shift is controlled by spin S . If the radio frequency is applied at a frequency δ (measured in Hz)

away from the transition frequency of spin *I* when spin *S* (the control spin) is in state 0, and ν_1 is the maximum radio-frequency field strength (also in Hz), then the differential Berry phase shift

$$\Delta\gamma = \gamma_1 - \gamma_0 = \pm \pi \left[\frac{\delta + J}{\sqrt{(\delta + J)^2 + \nu_1^2}} - \frac{\delta}{\sqrt{\delta^2 + \nu_1^2}} \right]$$

depends only on δ , ν_1 and J ; it is independent of how the process is carried out as long as it is slow enough to be adiabatic, but rapid compared with the decoherence times (T_1 and T_2).

In addition to the geometric phases, there will also be additional dynamic phases, which do depend on experimental details. In principle these could be calculated and corrected for, but this is not practical as a result of B_1 inhomogeneity. The radio-frequency field strength will vary over the sample, and so different nuclei will acquire different dynamic phases; averaging over the sample will result in extensive dephasing. This can be overcome using a conventional spin echo approach: the pulse sequence is applied twice, with the second application surrounded by a pair of 180° pulses applied to spin *I*. This has the effect of completely refocusing the dynamic phase, and thus refocusing any inhomogeneity in it. We note that this approach will only be successful if the dynamic phase terms are the same during the two halves of the spin echo, and thus it is important that any variation in these terms occurs on a timescale that is long compared with the echo time. In our experiments minor variations in the dynamic phase arising from the effects of molecular diffusion within the slightly inhomogeneous B_1 field are visible as a small loss in signal intensity. Similarly it is important to ensure that refocusing pulses are applied reliably, which is relatively simple within NMR. These issues must be considered when seeking to apply this approach with other experimental techniques.

This procedure would also cancel out the geometric phase, but this can be sidestepped by performing the radio-frequency phase sweep in the opposite direction, thus negating the geometric phase, so that the two geometric terms add together while the dynamic phases cancel out. Cancellation of dynamic phases arising from the natural evolution of spin *S*, could also be achieved by incorporating the sequence within another spin echo, involving 180° pulses applied to *S*. Similarly, in more complex spin systems, contributions to the geometric phase which depend on the states of other spins can be cancelled by the judicious application of further spin echoes. It might seem that this approach would require an exponentially large number of refocusing pulses, but this is not true, as nuclear spin-spin coupling is a local effect, so that couplings between distant nuclei can be safely neglected. It should also be possible to use efficient refocusing schemes¹⁷ based on Hadamard matrices, thus allowing the number of refocusing pulses to be further reduced.

In order to measure the sizes of γ_0 and γ_1 it is convenient to apply the Berry phase shifts to a spin *I* in a coherent superposition of states, created by an initial 90° pulse. The pulse sequence (Fig. 1)

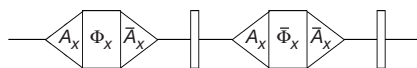


Figure 1 Pulse sequence used to demonstrate controlled Berry phases. Triangles indicate adiabatic radio-frequency amplitude sweeps, from 0 to ν_1 (A_x) or from ν_1 to 0 ($\bar{A}_x$); rectangles indicate slow rotations of the radio-frequency phase at constant amplitude. The phase rotation runs from 0° to 360° (Φ_x) or from 360° to 0° ($\bar{\Phi}_x$). Narrow rectangles correspond to hard 180° pulses. As the absolute phase of an NMR signal is undefined, it is essential to obtain a reference signal against which experimental phases can be measured. The simplest approach is to use the signal from a single 90° pulse, and a more subtle approach is to use this pulse sequence with the $\bar{\Phi}_x$ sequence replaced by Φ_x . In principle these should give the same result, but in practice minor differences are seen as a result of radio-frequency inhomogeneity and the effects of the long radio-frequency pulses on the NMR probe and pre-amplifier.

then generates Berry phases that are determined by examining the final phases of the magnetization. As the two states of spin *I* acquire equal and opposite phases, and the pulse sequence contains two separate periods in which phase shifts are generated, the total phase change observed is 4γ , with a maximum value of 720° . A range of controlled Berry phases can be generated by choosing appropriate values of δ and ν_1 (J is fixed by the chemical system). For a given value of δ/J the controlled phase will rise and then fall as ν_1 is increased. This approach will be most robust when the desired $\Delta\gamma$ occurs at the maximum of this curve, as the dependence on the size of δ is then reduced to second order. For the particular case of a controlled π shift, the basis of the controlled-NOT gate¹⁴, this occurs at $\delta = 1.058J$ and $\nu_1 = 2.112J$.

NMR experiments were performed at 25°C using a homebuilt 500 MHz (^1H frequency) NMR spectrometer at the OCMS, with two power ranges for ^1H pulses. Hard pulses were applied using high power ($\nu_1 < 25.8\text{ kHz}$), while adiabatic sweeps were performed using low power ($\nu_1 < 774\text{ Hz}$). Radio-frequency amplitude and phase-calibration tables (available on most modern NMR spectrometers) permit the radio-frequency power level to be varied in a phase-coherent manner. The sample was prepared by dissolving 100 mg of 99% ^{13}C -labelled CHCl_3 in 0.2 ml of 99.96% CDCl_3 , and placing this in a Shigemi microtube. The single ^1H nucleus was used as spin *I*, while the ^{13}C nucleus was used as spin *S*; for this system $J_{\text{IS}} = 209.2\text{ Hz}$. Spin-spin relaxation times (measured using Carr–Purcell–Meiboom–Gill sequences and averaging over the two components of each doublet) were 3.9 s for ^1H and 0.3 s for ^{13}C ; the spin–lattice relaxation times (measured by inversion-recovery) were 7.6 s for ^1H and 25.3 s for ^{13}C . Experiments were performed with $\delta = 221.3\text{ Hz}$, and ν_1 was varied between zero and its maximum value. Amplitude and phase sweeps were implemented using 200 linear steps of $100\text{ }\mu\text{s}$, giving a total pulse sequence length of about 120 ms. The phases of the two ^1H resonances were determined by fitting the free induction decay using home-written software. Reference phases were obtained as described in Fig. 1 (legend). The results are shown in Fig. 2; clearly the measured phases lie close to the theoretically predicted values. The controlled Berry phase rises smoothly to a broad maximum at 180° , and then slowly falls back towards zero. In order to investigate the effects of a breakdown in the adiabatic criterion, some measurements were repeated with faster sweeps. With a sweep step size of $50\text{ }\mu\text{s}$ (data not shown) the

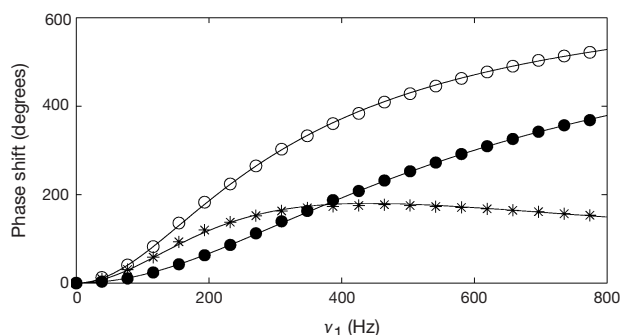


Figure 2 Experimental values for the Berry phases γ_0 , γ_1 , and the controlled Berry phase difference, $\Delta\gamma$, as a function of the maximum radio-frequency field strength ν_1 . Experimental points are shown by filled circles (γ_0), empty circles (γ_1) and stars ($\Delta\gamma$); theoretical values are shown as smooth curves. Variability in the experimental points (estimated by repetition) was about $\pm 2^\circ$; in a few cases the deviation of the measured data points from their theoretical values was greater than this, indicating the existence of as yet unidentified systematic errors. The signal strength observed after a phase gate was about 90% of that observed without a phase gate. This signal loss of about 10% is too great to arise simply from relaxation; more detailed experiments (data not shown) suggest that the main source of signal loss is the effect of diffusion. When considering the overall fidelity of the gate it is also necessary to include effects arising from spin *S*; these are dominated by the relatively rapid spin–spin (T_2) relaxation of the ^{13}C nucleus.

results were similar to, but not quite as good as, those obtained with the slower sweep. Below 50 μs the loss of adiabaticity is severe and major distortions are observed. The step size should not be increased too far beyond the adiabatic threshold, as this will increase the effects of decoherence.

The conditional Berry phase gate demonstrated here depends only on the geometry of the path, and is completely independent of how the motion is performed as long as it is adiabatic; hence this kind of computation may be called geometric quantum computation. While this approach has no particular advantage over more conventional methods in NMR quantum computation, the basis idea is general, and could be applied in other implementations. Some of the methods described here have been partly covered in previous theory papers (such as refs 18 and 19), but there have been no previous experimental demonstrations. This new approach to quantum gates may become important, as it is naturally resilient to certain types of errors. In particular, suppose that the qubit, in addition to the circular motion which implements the geometric phase, also undergoes a random motion about its path due to some unwanted interaction with the environment. Such noisy motion leaves the total area approximately unchanged (although it changes details of the path), and so will not be reflected in the final Berry phase. Geometric phases thus offer the potential of performing quantum computations in a manner which is naturally tolerant of some types of fault. Further generalizations to non-abelian Berry phases, if implemented, may open entirely new possibilities for robust quantum information processing^{20,21}. $\square$

Received 22 July; accepted 14 December 1999.

- Shor, P. W. Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. *SIAM Rev.* **41**, 303–332 (1999).
- Barenco, A., Deutsch, D., Ekert, A. & Jozsa, R. Conditional quantum dynamics and logic gates. *Phys. Rev. Lett.* **74**, 4083–4086 (1995).
- Berry, M. V. Quantal phase factors accompanying adiabatic changes. *Proc. R. Soc. Lond. A* **392**, 45–57 (1984).
- Shapere, A. & Wilczek, F. *Geometric Phases in Physics* (World Scientific, Singapore, 1989).
- Cory, D. G., Fahmy, A. F. & Havel, T. F. Ensemble quantum computing by NMR spectroscopy. *Proc. Natl Acad. Sci. USA* **94**, 1634–1639 (1997).
- Gershenfeld, N. A. & Chuang, I. L. Bulk spin-resonance quantum computation. *Science* **275**, 350–356 (1997).
- Jones, J. A. & Mosca, M. Implementation of a quantum algorithm on a nuclear magnetic resonance quantum computer. *J. Chem. Phys.* **109**, 1648–1653 (1998).
- Chuang, I. L., Gershenfeld, N. & Kubinec, M. Experimental implementation of fast quantum searching. *Phys. Rev. Lett.* **80**, 3408–3411 (1998).
- Cory, D. G., Price, M. D. & Havel, T. F. Nuclear magnetic resonance spectroscopy: An experimentally accessible paradigm for quantum computing. *Physica D* **120**, 82–101 (1998).
- Suter, D., Chingas, G., Harris, R. & Pines, A. Berry's phase in magnetic resonance. *Mol. Phys.* **61**, 1327–1340 (1987).
- Goldman, M., Fleury, V. & Guéron, M. NMR frequency shift under sample spinning. *J. Magn. Reson. A* **118**, 11–20 (1996).
- Jones, J. A. & Pines, A. Geometric dephasing in zero-field magnetic resonance. *J. Chem. Phys.* **106**, 3007–3016 (1997).
- Deutsch, D., Barenco, A. & Ekert, A. Universality in quantum computation. *Proc. R. Soc. Lond. A* **449**, 669–677 (1995).
- Jones, J. A., Hansen, R. H. & Mosca, M. Quantum logic gates and nuclear magnetic resonance pulse sequences. *J. Magn. Reson.* **135**, 353–360 (1998).
- Ernst, R. R., Bodenhausen, G. & Wokaun, A. *Principles of Nuclear Magnetic Resonance in One and Two Dimensions* (Clarendon, Oxford, 1987).
- Sørensen, O. W., Eich, G. W., Levitt, M. H., Bodenhausen, G. & Ernst, R. R. Product operator-formalism for the description of NMR pulse experiments. *Prog. Nucl. Magn. Reson. Spectrosc.* **16**, 163–192 (1983).
- Jones, J. A. & Knill, E. Efficient refocussing of one spin and two spin interactions for NMR quantum computation. *J. Magn. Reson.* **141**, 322–325 (1999).
- Pellizzari, T., Gardiner, S. A., Cirac, J. I. & Zoller, P. Decoherence, continuous observation, and quantum computing: a cavity QED model. *Phys. Rev. Lett.* **75**, 3788–3791 (1995).
- Averin, D. V. Adiabatic quantum computation with Cooper pairs. *Solid State Commun.* **105**, 659–664 (1998).
- Kitaev, A. Y. Fault tolerant quantum computation with anyons. Preprint <http://arxiv.org/quant-ph/9707021>.
- Preskill, J. Fault tolerant quantum computation. Preprint <http://arxiv.org/quant-ph/9712048>.

Acknowledgements

We thank N. Soffe for helpful discussions. J.A.J. and A.E. thank the Royal Society of London and Starlab (Riverland NV) for financial support.

Correspondence and requests for materials should be addressed to J.A.J. (e-mail: jonathan.jones@qubit.org).

Reduction in the surface energy of liquid interfaces at short length scales

C. Fradin*, A. Braslau*, D. Luzet*, D. Smilgies†, M. Alba*, N. Boudet‡, K. Mecke‡ & J. Daillant*§

* Service de Physique de l'Etat Condensé, CEA Saclay, F-91191 Gif-sur-Yvette Cedex, France

† European Synchrotron Radiation Facility, BP 220, F-38043 Grenoble Cedex, France

‡ Theoretische Physik, Bergische Universität Wuppertal, D-42097 Wuppertal, Germany

§ LURE, Centre Universitaire Paris-sud, bâtiment 209D, PB 34, F-91898 Orsay Cedex, France

Liquid–vapour interfaces, particularly those involving water, are common in both natural and artificial environments. They were first described as regions of continuous variation of density¹, caused by density fluctuations within the bulk phases^{2–4}. In contrast, the more recent capillary-wave model^{5,6} assumes a step-like local density profile across the liquid–vapour interface, whose width is the result of the propagation of thermally excited capillary waves. The model has been validated for length scales of tenths of micrometres and larger^{7,8}, but the structure of liquid surfaces on submicrometre length scales—where the capillary theory is expected to break down—remains poorly understood. Here we report grazing-incidence X-ray scattering experiments that allow for a complete determination of the free surface structure and surface energy for water and a range of organic liquids. We observe a large decrease of up to 75% in the surface energy of submicrometre waves that cannot be explained by capillary theory, but is in accord with the effects arising from the non-locality of attractive intermolecule interactions as predicted by a recent density functional theory⁹. Our data, and the results of comparable measurements on liquid solutions, metallic alloys, surfactants, lipids and wetting films should thus provide a stringent test for any new theories that attempt to describe the structure of liquid interfaces with nanometre-scale resolution.

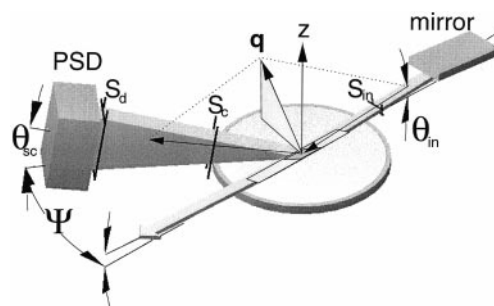


Figure 1 Schematic view of the experiment. The Teflon trough (inner diameter is 330 mm) is mounted on an active antivibration system under a helium atmosphere saturated with the vapour of the liquid under study. The monochromatic incident beam is first extracted from the polychromatic beam of the undulator source using a two-crystal diamond (111) monochromator. Higher harmonic light is eliminated using two platinum-coated glass mirrors, also used to fix the grazing angle of incidence θ_{in} . The incident and scattered beams travel through vacuum paths and the scattered signal is collected in a vertically mounted position sensitive detector (PSD). The size of the incident beam is fixed by a $265 \mu\text{m} \times 250 \mu\text{m}$ slit, S_n . The horizontal resolution of the experiment is fixed by the slits S_c , $380 \mu\text{m}$ in width and S_d , $545 \mu\text{m}$ in width, placed at 213 mm and 844 mm from the sample respectively, as well as by the illuminated area seen by the detector (dark grey parallelogram in centre of circle). The scattered beam is defined by the angles Ψ and θ_{sc} . q is the wavevector transfer.

The liquid surface structure, described by the height correlation spectrum $\langle z(q)z(-q) \rangle \propto k_B T / \gamma q^2$, where γ is the surface tension and $2\pi/q$ is the wavelength of the capillary excitation of amplitude z and wavevector q , is determined by the surface energy associated with the deformation modes. Here k_B is Boltzmann's constant and T the

temperature. Subtle effects beyond the capillary theory can be included by considering an effective hamiltonian or wavevector-dependent surface energy $\gamma(q)$ (refs 9–12). For example, for large wavevector moduli of the order of the inverse of the local interface thickness, the distortion of the density profile caused by the interface curvature is no longer negligible, and this leads to a larger effective surface tension^{13,14}, that is, a less rough interface. A larger effective surface tension is also predicted at small length scales considering either the free energy carried by the capillary waves themselves^{5,11}, or the renormalization by the short wavelength fluctuations of the surface tension for the longer wavelength fluctuations¹⁵. The experimental verification of such effects has been long sought after^{16–18}, but no experimental technique has been able to provide information resolved at large enough wavevectors.

For our experiments we used a 8.27-keV (wavelength $\lambda = 0.150$ nm, wavevector $k_0 = 2\pi/\lambda = 4.19 \times 10^{10} \text{ m}^{-1}$) X-ray beam at the Troika II beamline of the European Synchrotron Radiation Facility (ESRF), see Fig. 1. Within the distorted-wave Born approximation^{19–21} which is currently the most accurate theory for the analysis of X-ray surface scattering experiments, the scattering cross-section (power radiated per unit solid angle in the scattering direction per unit incident flux) can be approximated for large in-plane wavevector transfers $q_{||}$ (see Fig. 1) by:

$$\left(\frac{d\sigma}{d\Omega} \right) \approx A \frac{k_0^4 \theta_c^4}{16\pi^2} |t^{\text{in}}|^2 |t^{\text{sc}}|^2 \left[\frac{k_B T}{\gamma q_{||}^2} \left(\frac{q_{||}}{q_{\text{max}}} \right)^\eta + \frac{k_B T \kappa_T}{2\text{Im}(q_z^t)} \right] \quad (1)$$

Here $\eta = k_B T |q_z^t|^2 / (2\pi\gamma)$ is a small exponent less than 0.16 in these experiments. The first contribution within the square brackets is that of the surface fluctuations^{8,22,23} (capillary waves, $\gamma = 72.8 \text{ mN m}^{-1}$ for water at room temperature), and the second one that of propagating or non-propagating density fluctuations (κ_T is the isothermal compressibility, equal to $4.58 \times 10^{-10} \text{ Pa}^{-1}$ for water), both of which contribute to the scattered signal. The surface contribution is proportional to the height fluctuation spectrum for small values of the vertical component of the wavevector transfer q_z and the bulk contribution, which does not depend on $q_{||}$, will inevitably dominate the surface contribution for large $q_{||}$ values. The imaginary part of the vertical component of the transmitted wavevector transfer in the liquid $\text{Im}(q_z^t)$ fixes the effective penetration length of the X-ray radiation. The largest wavevector in the capillary wave spectrum q_{max} is of the order of an inverse molecular diameter. A is the illuminated area on the surface. t^{in} and t^{sc} are the Fresnel transmission coefficients for the incident and scattering angle, respectively. The critical angle for total external reflection, which occurs for X-rays because the refractive index of matter is less than 1, is $\theta_c = 2.60 \text{ mrad}$ for water at 8.27 keV. Depending on whether the grazing angle of incidence is below or above the critical angle for total external reflection, the effective penetration length in the experiment varies from less than 6 nm whatever the scattering angle (evanescent wave) to more than $10 \mu\text{m}$, allowing control of the relative weight of the surface and bulk terms. Note that minimizing the penetration depth is essential in order to measure the very small $q_{||}$ surface signal (less than 10^{-11} of the incident intensity). Using equation (1), the scattered intensity is obtained by numerically integrating the scattering cross-section over the detector angular acceptance.

We first measured the scattering from the surface of water at room temperature. The ultra-pure de-ionized water was produced by a Milli-Q 185+ Millipore system. The measurements were performed under a vapour-saturated helium atmosphere in order to minimize the background, which was recorded by lowering the trough and repeating the measure²⁴, and then subtracted from the measurements. This procedure yields a precise estimation of the background in the region of interest $10 \text{ mrad} \leq \theta_{\text{sc}} \leq 0.1 \text{ rad}$ under total external reflection conditions. The measurements were repeated several times, with the same or different samples, in order to rule out any

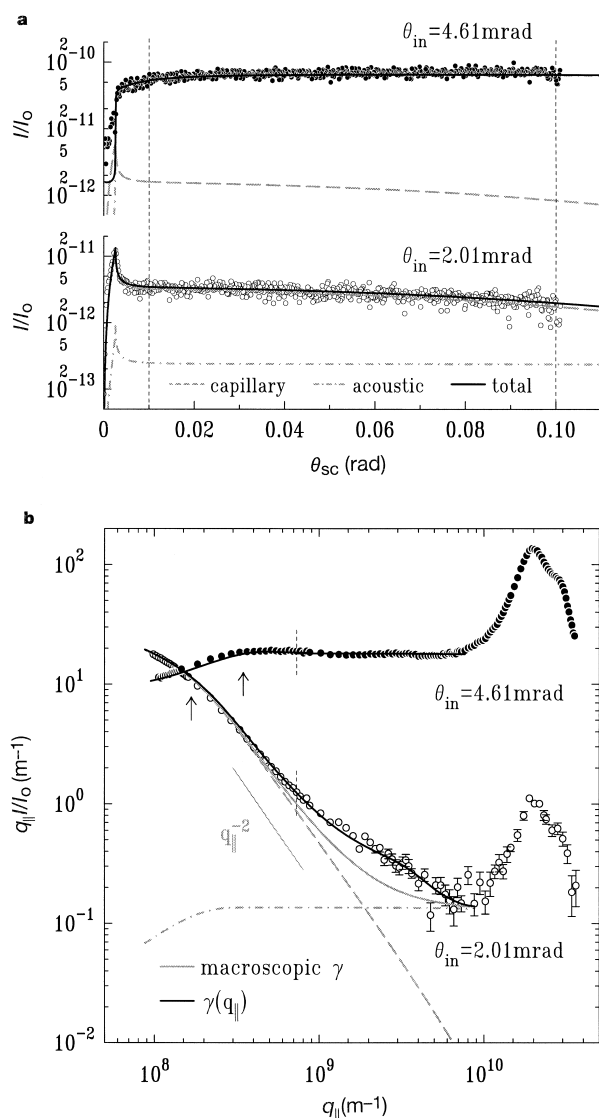


Figure 2 Scattering at the liquid water–vapour interface. Two values of the grazing angle of incidence are shown: $\theta_{\text{in}} = 4.61 \text{ mrad}$ above the critical angle for total external reflection θ_c (black circles, bulk scattering dominant); and $\theta_{\text{in}} = 2.01 \text{ mrad} < \theta_c$ (empty circles, capillary wave contribution dominant for all but high $q_{||}$ values). The continuous grey lines are the result of calculations using equation (1), split into capillary-wave contribution using the macroscopic surface tension γ (grey long-dashed lines), and acoustic-wave contribution (grey short-dashed lines). The continuous black lines have been calculated using $\gamma(q_{||})$ given by the theory³. There is no adjustable parameter in any of those calculations. **a**, Signal recorded in the detector (circles) for $q_{||} = 7.31 \times 10^8 \text{ m}^{-1}$ ($\psi = 0.0175 \text{ rad}$, indicated by vertical dotted lines in **b**). The peak at $\theta_{\text{sc}} = \theta_c$ is known as the Yoneda peak²⁰. The vertical dotted lines indicate the range of integration of the signal in θ_{sc} for the curves presented in **b**. **b**, Integrated signal in the PSD (see Fig. 1; between $10 \text{ mrad} \leq \theta_{\text{sc}} \leq 0.1 \text{ rad}$) as a function of the horizontal component of the wavevector transfer $q_{||}$. The experimental signal has been multiplied by $q_{||}$ in order to compensate approximately for resolution effects and to obtain curves proportional to the scattering cross-section or to the fluctuation spectra (at least above a resolution cut-off indicated by arrows). Note the peak at $q_{||} \approx 2 \times 10^{10} \text{ m}^{-1}$ giving the short-range structure of the nearest neighbours in water. The two peaks for $\theta_{\text{in}} = 4.61 \text{ mrad}$ and 2.01 mrad superpose precisely when we take the penetration depth difference into account. A line parallel to q^{-2} is indicated as a guide for the eye.

time-dependent contamination. All the measurements agreed within the error bars, and we present here representative measurements made for grazing angles of incidence above ($\theta_{\text{in}} = 4.61$ mrad) and below ($\theta_{\text{in}} = 2.01$ mrad) the critical angle for water at 8.27 keV. The z -dependence of the scattered intensity in the detector is shown in Fig. 2a, while the resulting integrated intensity ($10 \text{ mrad} \leq \theta_{\text{sc}} \leq 0.1$ rad) is given as a function of $q_{\parallel}$ in Fig. 2b. We note (Fig. 2a) the accuracy of the theoretical description of the scattered intensity without any adjustable parameter, as well as the decomposition in bulk and surface fluctuations.

Figure 2b displays the first, to our knowledge, complete measurement of the structure of a liquid interface down to molecular scale and gives a striking picture of the very nature of a liquid surface. Starting from high $q_{\parallel}$ values, that is, at local scales, we observe the static peak corresponding to the short-range order of the nearest neighbours. The same peak is observed in the bulk if the penetration length is increased. For the surface-sensitive measurement, and at larger length scales (smaller $q_{\parallel}$ values), we observe the characteristic divergence in the spectrum. In contrast, there is no divergence in the low- $q_{\parallel}$ end of the bulk spectrum which is well described by the white spectrum of density fluctuations. More precisely, the surface-sensitive experimental data are correctly described by the simple capillary wave spectrum only up to $q_{\parallel} \approx 7 \times 10^8 \text{ m}^{-1}$. Beyond this value, the acoustic wave contribution within the penetration depth becomes important, but the experimental scattering is still significantly larger than calculated for $7 \times 10^8 \text{ m}^{-1} \leq q_{\parallel} \leq 10^{10} \text{ m}^{-1}$. This additional scattering is not seen in the data taken for a grazing angle of incidence larger than θ_c , except for very small variations of the order of the surface signal (two orders of magnitude weaker than the bulk scattering) and must be due to the surface region. Attempts to fit the data by assuming an excess amplitude of acoustic waves (that is, a larger compressibility) in the near-surface region failed, and, more generally, density fluctuation effects limited to the near-surface region do not yield the correct order of magnitude. It should also be noted that the same excess scattering effect was observed for very different liquids: ethylene glycol, carbon tetrachloride and hexadecane, eliminating the possibility of a water-specific effect such as orientation fluctuations. What we observe must therefore be a universal effect affecting the short-scale surface morphology. The surface is found to be rougher than expected from the simple capillary model, contrary to what is predicted by the phenom-

ological capillary wave theories discussed above. A rougher surface at small length scales is equivalent to a smaller surface energy $\gamma(q_{\parallel})$, which can be estimated from the measurements: by subtracting first the known bulk contribution (equation (1)), the enhanced surface modes will be inversely proportional to $\gamma(q_{\parallel})$ (equation (1)). $\gamma(q_{\parallel})/\gamma$ can therefore be obtained by dividing the scattered intensity calculated for a simple $q_{\parallel}^{-2}$ spectrum by the experimental surface scattering (Fig. 3).

An explanation for the surprising lowering of the surface energy can be found in a recent density functional theory^{9,25} taking into account the long-range power-law decay of the dispersion forces always existing between molecules. Indeed, the lack of compensation of molecular interactions at the surface, which is the origin of surface tension, is reduced for a corrugated interfacial configuration having a wavelength shorter than the interaction range, of the order of several nanometres for the van der Waals dipolar interaction, implying a smaller surface energy at this scale. $\gamma(q_{\parallel})$ should first decrease from its value at $q_{\parallel} = 0$ due to the effect of attractive long-range forces⁹, reach a minimum and then increase proportional to $q_{\parallel}^2$ at large $q_{\parallel}$ owing to the distortion of density profile when the surface is bent.

The effective theoretical surface tension

$$\gamma(q_{\parallel}) = \int \int_{-\infty}^{\infty} dz dz' W(q_{\parallel}, z - z'; \tau_0) \rho(q_{\parallel}, z; \xi, C_H) \rho(q_{\parallel}, z'; \xi, C_H) \quad (2)$$

can be written as an integral over a function $W(q_{\parallel}, z - z')$ describing the molecular interaction potential and the $q_{\parallel}$ -dependent density profiles $\rho(q_{\parallel}, z)$, where z and z' denote the distances perpendicular to the surface⁹. Assuming standard expressions for dispersion forces and density profile, the surface tension $\gamma(q_{\parallel})$ depends only on the size r_0 of the molecules (hard-core diameter), on the correlation length ξ , and on the amplitude C_H of the curvature corrections of the density profile for thermally excited capillary modes. Whereas values for r_0 and ξ are experimentally accessible, for the parameter C_H only the limiting values $0 < C_H < 1$ can be given. The effective theoretical surface tension $\gamma(q_{\parallel})$, obtained using reasonable parameter values for water, is given in Fig. 3. The shape of the measured curve is perfectly described by the calculations where no fitting parameter is used. A negative correction to the short-scale surface tension was also found in recent molecular dynamics simulations²⁶, and a similar effect of the dispersion forces was recently observed in the bulk structure factor of rare gases²⁷. This may call for a re-examination of small-scale interfacial processes involving fluids in physics, chemistry or biology because any interface deformation at these scales should thus be easier than expected. □

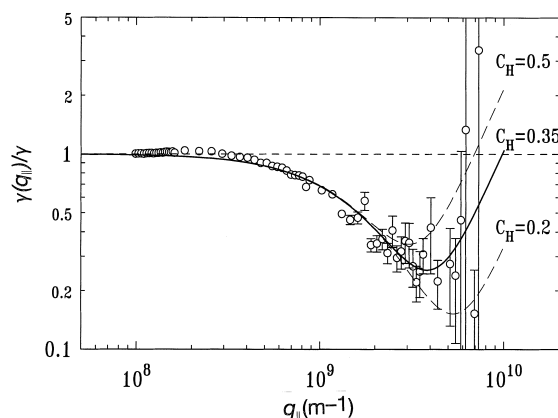


Figure 3 Wavevector dependent surface energy normalized to the macroscopic surface tension $\gamma(q_{\parallel})/\gamma$. Circles, values obtained by normalizing the capillary-wave contribution to the scattering to that expected for a fluctuating surface having the macroscopic surface tension $\gamma = \gamma(q_{\parallel} = 0)$ of water at all $q_{\parallel}$ values. The lines give the result of the density functional theory given in ref. 9 for $C_H = 0.35$ (continuous line), $C_H = 0.2$ (lower dashed line) and $C_H = 0.5$ (upper dashed line). The decrease of the surface energy $\gamma(q_{\parallel})$ does not depend sensitively on the molecular size r_0 and the correlation length ξ for values within a reasonable physical range.

Received 13 May 1999; accepted 14 January 2000.

- van der Waals, J. D. The thermodynamic theory of capillarity under the hypothesis of a continuous variation of density. *Verh. Konink. Akad. Wet. Amsterdam (Sect. 1)* **1** (1893). English translation by Rowlinson, J. S. *J. Stat. Phys.* **20**, 197–244 (1979).
- Rowlinson, J. S. & Widom, B. *Molecular Theory of Capillarity* Vol. 8 (International Series of Monographs on Chemistry, Oxford Univ. Press, 1982).
- Cahn, J. W. & Hilliard, J. E. Free energy of a nonuniform system. i. Interfacial free energy. *J. Chem. Phys.* **28**, 258–267 (1958).
- Fisk, S. & Widom, B. Structure and free energy of the interface between fluid phases in equilibrium near the critical point. *J. Chem. Phys.* **50**, 3219–3227 (1960).
- Buff, F. P., Lovett, R. A. & Stillinger, R. H. Interfacial density profile for fluids in the critical region. *Phys. Rev. Lett.* **15**, 621–623 (1965).
- Gelfand, M. P. & Fisher, M. E. Finite-size effects in fluid interfaces. *Physica A* **166**, 1–74 (1990).
- Loudon, R. in *Modern Problems in Condensed Matter Sciences* (eds Agranovich, V. M. & Loudon, R.) Vol. 9, 589–638 (North-Holland Physics, Amsterdam, 1984).
- Sanyal, M. K., Sinha, S. K., Huang, K. G. & Ocko, B. M. X-ray scattering study of capillary-wave fluctuations at a liquid surface. *Phys. Rev. Lett.* **66**, 628–631 (1991).
- Mecke, K. & Dietrich, S. Effective hamiltonian for liquid-vapor interfaces. *Phys. Rev. E* **59**, 6766–6784 (1999).
- Weeks, J. D. Structure and thermodynamics of the liquid-vapor interface. *J. Chem. Phys.* **67**, 3106–3121 (1977).
- Kayser, R. F. Effect of capillary waves on surface tension. *Phys. Rev. A* **33**, 1948–1956 (1986).
- Sengers, J. V. & van Leeuwen, J. M. J. Capillary waves of a vapor-liquid interface near the critical temperature. *Phys. Rev. A* **39**, 6346–6355 (1989).

13. Tolman, R. C. J. The effect of droplet size on surface tension. *J. Chem. Phys.* **17**, 333–337 (1949).
14. van Giessen, A. E., Blokhuis, E. M. & Bukman, D. J. Mean field curvature corrections to the surface tension. *J. Chem. Phys.* **108**, 1148–1156 (1998).
15. Meunier, J. Liquid interfaces: role of the fluctuations and analysis of ellipsometry and reflectivity measurements. *J. Phys.* **48**, 1819–1831 (1987).
16. Braslau, A. *et al.* Surface roughness of water measured by x-ray reflectivity. *Phys. Rev. Lett.* **54**, 114–117 (1985).
17. Rice, S. A. X-ray reflection from liquids. *Nature* **316**, 108 (1985).
18. Daillant, J., Bosio, L., Harzallah, B. & Benattar, J. J. Structural properties and elasticity of amphiphilics on water. *J. Phys. II* **1**, 149–170 (1991).
19. Sinha, S. K., Sirota, E. B. & Garoff, S. X-ray and neutron scattering from rough surfaces. *Phys. Rev. B* **38**, 2297–2311 (1988).
20. Daillant, J. & Belorgey, O. Surface scattering of x-rays in thin films. i. Theoretical treatment. *J. Chem. Phys.* **97**, 5824–5836 (1992).
21. Dietrich, S. & Haase, A. Scattering of x-rays and neutrons at interfaces. *Phys. Rep.* **260**, 1–138 (1995).
22. Fukuto, M. *et al.* X-ray measurements of noncapillary spatial fluctuations from a liquid surface. *Phys. Rev. Lett.* **81**, 3455–3458 (1998).
23. Tostmann, H. *et al.* Surface structure of liquid metals and the effect of capillary waves: x-ray studies on liquid indium. *Phys. Rev. B* **59**, 783–791 (1999).
24. Gourier, C. *et al.* Bending energy of amphiphilic films at the nanometer scale. *Phys. Rev. Lett.* **78**, 3157–3160 (1997).
25. Napiórkowski, M. & Dietrich, S. Structure of the effective hamiltonian for liquid-vapor interfaces. *Phys. Rev. A* **47**, 1836–1849 (1993).
26. Stecki, J. Extended capillary wave theory and the ellipsometric coefficient. *J. Chem. Phys.* **109**, 5002–5007 (1998).
27. Magli, R., Barocchi, F., Chieux, P. & Fontana, R. Experimental determination of the long-range potential of argon pairs by means of small-angle neutron scattering. *Phys. Rev. Lett.* **77**, 846–848 (1996).

Acknowledgements

We thank C. Blot, P. Feder and H. Gleyzolle for their help.

Correspondence and requests for materials should be addressed to J.D.
(e-mail: daillant@spec.saclay.cea.fr or daillant@lure.u-psud.fr).

Electrically induced structure formation and pattern transfer

Erik Schäffer*, Thomas Thurn-Albrecht†, Thomas P. Russell‡ & Ullrich Steiner*‡

* Fakultät für Physik, Universität Konstanz, 78457 Konstanz, Germany

† Polymer Science and Engineering Department, University of Massachusetts at Amherst, Amherst, Massachusetts 01003, USA

‡ Department of Polymer Chemistry, University of Groningen, Nijenborgh 4, 9747 AG Groningen, The Netherlands

The wavelength of light represents a fundamental technological barrier¹ to the production of increasingly smaller features on integrated circuits. New technologies that allow the replication of patterns on scales less than 100 nm need to be developed if increases in computing power are to continue at the present rate². Here we report a simple electrostatic technique that creates and replicates lateral structures in polymer films on a sub-micrometre length scale. Our method is based on the fact that dielectric media experience a force in an electric field gradient³. Strong field gradients can produce forces that overcome the surface tension in thin liquid films, inducing an instability that features a characteristic hexagonal order. In our experiments, pattern formation takes place in polymer films at elevated temperatures, and is fixed by cooling the sample to room temperature. The application of a laterally varying electric field causes the instability to be focused in the direction of the highest electric field. This results in the replication of a topographically structured electrode. We report patterns with lateral dimensions of 140 nm, but the extension of the technique to pattern replication on scales smaller than 100 nm seems feasible.

The interaction of an electric field with matter is well understood³. Common experiments demonstrate the forces on dielectric materials in electric fields, such as a dielectric liquid which is drawn into the air gap of a plate capacitor when a voltage is applied across the plates. As electrostatic interactions are relatively strong and long-range, they can be used to control structures on length scales which are difficult to manipulate in any other manner. One challenge in this context is the creation of lateral features below the diffraction limit of light. Various techniques for the manufacture of lithographic layers with submicrometre structures have been proposed, such as particle lithography, lithography using scanning near-field microscopes, and printing or embossing structures onto surfaces². Here we describe a particularly simple addition to this collection of submicrometre replication techniques: the electrostatic transfer of a master pattern into a polymer film. While electric fields have been used before to orient the microphases in block copolymer melts^{4–8}, the work presented here focuses on the use of electric fields to control and organize polymers laterally to replicate a master pattern.

This work has two main aspects. First, we show that a thin liquid film can be destabilized by strong electric fields, applied across the film by means of two capacitor plates. This instability leads to the formation of ordered lateral structures. We also present a theoretical model describing the destabilization process. Second, we make use of this effect for electrostatic lithography: one of the capacitor plates is used as a topographic master. Patterns down to lateral dimensions of 140 nm have been replicated.

A sketch of the experimental set-up is shown in Fig. 1a. Initially, a thin polymer film of thickness h was spin-coated from solution onto a highly polished silicon wafer serving as one of the electrodes. Subsequently, another silicon wafer was mounted as the opposing electrode at a distance d , leaving a thin air gap. The assembly was then heated above the glass transition temperature of the polymer (T_g), and a small voltage U (20–50 V) was applied. To assure the air gap, the top electrode has a small step. Using a wedge geometry, a range of d values could be achieved, while locally maintaining a nearly parallel electrode configuration. Air gaps varying from 100 nm to 1,000 nm were obtained this way. The voltage and the geometry of the capacitor device determine the electric field. The electrostatic driving force scales with the difference between the electric field in the polymer film, E_p , and the field in the air gap, E_a . E_p increases with decreasing values of d , and with increasing polymer film thickness h . The low applied voltage, combined with the small distance between the electrodes ($d < 1 \mu\text{m}$), leads to high electric fields E_p (10^7 – 10^8 V m^{-1}). During the annealing step, a small current

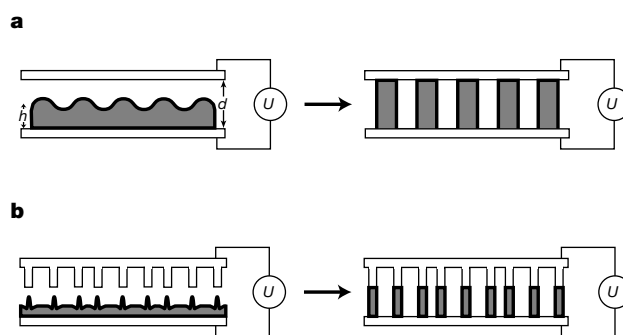


Figure 1 Schematic representation of the capacitor device. **a**, The electrostatic pressure acting at the polymer (grey)–air interface causes an instability in the film (left). This instability has a well defined wavelength. Eventually, polymer columns span the gap between the two electrodes (right). **b**, If the top electrode is replaced by a topographically structured electrode, the instability occurs first at the locations where the distance between the electrodes is smallest (left). This leads to replication of the electrode pattern (right).

was flowing through the device ($10\text{--}50\text{ mA cm}^{-2}$). At high electric fields, the current is caused by an ion conduction mechanism mediated by small impurity molecules in the polymer matrix⁹. After an annealing time which ranged from several minutes to a few hours, the polymer was immobilized by quenching below T_g , the top electrode was mechanically removed, and the morphology of the polymer film was investigated by optical microscopy and atomic force microscopy (AFM).

The results of typical experiments are shown in the optical micrographs of Fig. 2. Depending on the electric field strength and the annealing time, we observe either surface undulations (Fig. 2a) or a columnar structure where the liquid bridges the gap between the electrodes (Fig. 2b). As the two electrodes are not perfectly parallel, the electric field exhibits a small lateral variation and we usually observe both situations (Fig. 2a, b) on the same sample. In the case of the columnar phase, the electrode spacing can be measured by the AFM. We emphasize that the structures in Fig. 2 are observed only with the application of a very high electric field ($>10^7\text{ V m}^{-1}$). We observed no surface features in the absence of an applied field, in contrast to recent, intriguing results by Chou *et al.*^{10,11}. For significantly longer annealing times, dispersive van der Waals interactions will cause the breakup of the film¹². When the instability is caused by purely dispersive interactions, very different time scales of breakup of the film are observed and the film morphology is typically stochastic and not ordered. This was confirmed by a control experiment without applied voltage (not shown here), in which the film remained stable.

Different degrees of order are evident in Fig. 2. The ordering phenomenon originates from the repulsion of the equally charged undulation maxima and minima. While the film morphology in Fig. 2b exhibits only imperfect order, a more complete hexagonal packing is achieved in Fig. 2c. The main difference between the two images is the initial thickness h of the polymer film (93 and 193 nm, respectively). With otherwise similar parameters, this leads to a more dense lateral arrangement of the polymer columns and to an

increased column–column repulsion. As a consequence, improved hexagonal order is observed in Fig. 2c, compared to Fig. 2b.

While nearest-neighbour interactions lead to a hexagonal symmetry, second-order effects can be observed as well, as demonstrated for the nucleated instability in Fig. 2d. The locally higher value of h at the nucleation point leads to a higher electric field and an increased driving force. This causes a depletion of the nearest-neighbour undulations. The next-nearest neighbours, on the other hand, are again amplified. They form a rosette on a circle with a radius $r = 2\lambda$ and a circumference of $2\pi r \approx 12\lambda$, where λ is a characteristic wavelength. Beyond the circle of next-nearest neighbours, the instability decays with increasing distance. In the absence of nucleation effects, a similar argument also explains the hexagonal closest packing of the columns in Fig. 2c, where each maximum is surrounded by six neighbours, corresponding to a circle with radius λ and circumference $\sim 6\lambda$.

We also note the well defined lateral length scale of the surface undulations. The wavelength λ is a function of the electric field E_p , which varies inversely with the electrode spacing. The lateral structure dimensions, as well as the plateau height, are readily measured with the AFM, yielding λ as a function of electrode spacing d . In Fig. 3, λ is plotted as a function of the electric field in the polymer layer, E_p . For a given film thickness h , the characteristic lateral structure size exhibits a power-law dependence on the increasing electric field, corresponding to a decrease in the electrode spacing (top axis).

Theoretically, the origin of the film instability can be understood by considering the balance of forces that act at the polymer–air interface. The surface tension γ minimizes the area of the polymer–air surface, and stabilizes the homogeneous polymer film. The electric field, on the other hand, polarizes the dielectric. This results in an effective displacement charge density at the liquid–air interface, which destabilizes the film. When this balance of forces is written in terms of pressures, an electrostatic pressure p_{el} is opposed by the Laplace pressure. A local perturbation in the film thickness

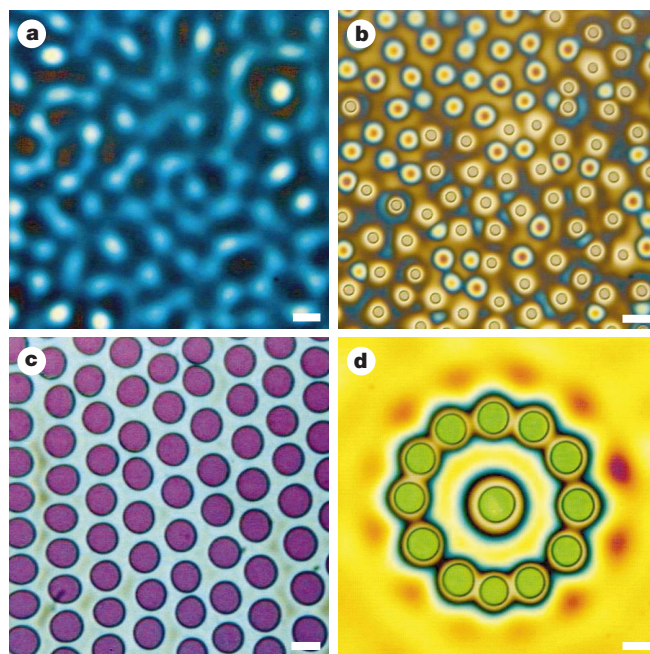


Figure 2 Optical micrographs of polystyrene films which have been exposed to an electric field. In **a** and **b**, a 93-nm-thick polystyrene film was annealed for 18 h at 170°C with an applied voltage $U = 50\text{ V}$. As the electrode spacing in **a** was larger than in **b**, corresponding to a lower electric field E_p , two different stages of the instability were observed, corresponding to Fig. 1a and b. In **c** and **d**, the thickness of the polystyrene film was doubled ($h = 193\text{ nm}$). This leads to a denser packing of the polymer columns and to an enhanced repulsion between the columns. As a result, we observe a more complete

hexagonal order in **c**. In **d** a second-order effect is observed for a nucleated instability (see text). Instead of a hexagonal symmetry which characterizes the nearest neighbours in **c**, a ring of 12 columns lies on a circle with a radius of 2λ . The scale bars correspond to $10\text{ }\mu\text{m}$ in **a** and **b** and to $5\text{ }\mu\text{m}$ in **c** and **d**. The colours arise from the interference of light, and correspond to the local thickness of the polymer structures (for example, in **d**, yellow corresponds to a film thickness of $\sim 200\text{ nm}$, green to $\sim 450\text{ nm}$).

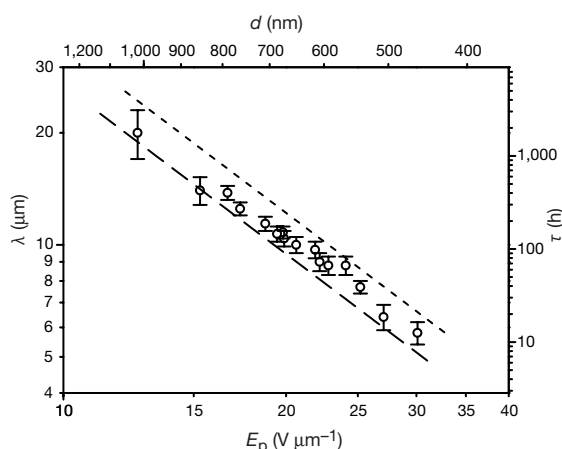


Figure 3 Variation of the lateral wavelength λ as a function of the electric field E_p in the polymer film. The open circles are experimental data for the set-up in Fig. 1a with a 93-nm-thick polystyrene film and an applied voltage of 30 V. The sample was annealed for 120 h at 170 °C. After removal of the top plate, λ and the electrode spacing d were measured by imaging the polymer columns with the AFM. Also plotted are the theoretical predictions from equation (1). The electrostatic pressure is computed as: $p_{el} = -f(\epsilon)E_p^2$. The precise functional dependence $f(\epsilon)$ of p_{el} on the dielectric constant ϵ of the polymer is not known. Two slightly different models yield the dotted and dashed lines which bound the experimental data. The right-hand axis indicates the predicted time constant τ for the pattern-formation process corresponding to the dashed line.

results in a pressure gradient which drives a flow of the dielectric liquid in the plane of the film. The liquid flow next to a solid surface is given by a formula of the Poiseuille type¹³ which, together with a mass-conservation equation, establishes a differential equation describing the temporal response of the liquid. A common approach to the investigation of the effect of external forces on a liquid film is a linear stability analysis. A small sinusoidal perturbation is applied to an otherwise flat film, and its response is calculated using a linearized version of the differential equation¹⁴. The resulting dispersion relation quantifies the decay, or amplification, of a given perturbation wavelength with time. The wavelength of the fastest amplified mode is given by:

$$\lambda_m = 2\pi \sqrt{\frac{2\gamma}{\frac{\partial p_{el}}{\partial h}}} \quad (1)$$

Here p_{el} is a function of the electric field and the dielectric constant of the polymer (Fig. 3). The dotted and dashed lines in Fig. 3 show λ_m as a function of the electric field for slightly different models of the electrostatic pressure p_{el} . A similar equation quantifies the characteristic time τ for the formation of the instability (right-hand axis in Fig. 3). We note that the two curves from equation (1) describe the experimental data well, in the absence of any fitting parameters. Additional data using other polymers, along with a detailed theoretical model, will be discussed elsewhere.

Whereas the topography of the polymer film occurs spontaneously, control of the lateral structure is achieved by laterally varying the electrode spacing d . To this end, the upper electrode is replaced by a topographically patterned master (Fig. 1b). As the electrostatic forces are strongest for smallest electrode spacings d , the time needed for the instability to form is much shorter for smaller values of d (Fig. 3, right hand axis). The emerging structure in the film is therefore focused towards the electrode structure which protrudes downwards towards the polymer film (Fig. 1b). This leads to a replication of the master electrode. Experimental results of this procedure are shown in Fig. 4. We note that the dielectric liquid is drawn up towards the electrode, forming a positive replica, as opposed to the negative patterns achieved by

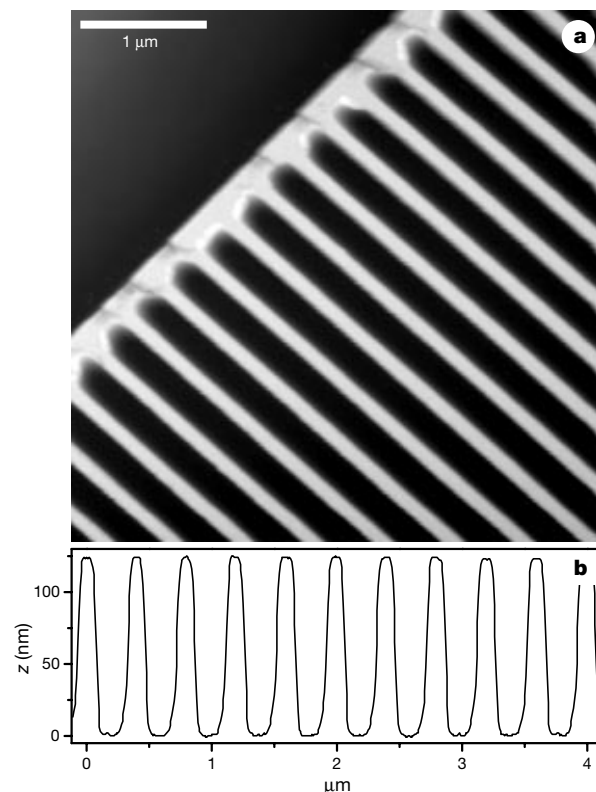


Figure 4 Electrostatic lithography corresponding to the schematic drawing in Fig. 1b. A patterned electrode was mounted facing a brominated polystyrene film ($h \approx 45$ nm). The device was annealed at 170 °C for 14 h. To ensure that no polymer remained on the master electrode after disassembly, the electrode was rendered unpolar by deposition of a self-assembled monolayer. The AFM image (**a**; tapping mode at 360 kHz) shows 140-nm-wide stripes (full-width at half-maximum) which replicate the silicon master electrode (200-nm stripes separated by 200-nm-wide and 170-nm-deep grooves). The cross-section (**b**) reveals a step height of 125 nm. The resolution in **b** is limited by the geometry of the AFM tip. The rounded-off corners at $z = 0$ nm indicate a finite contact angle of the brominated polystyrene on the silicon oxide surface. Apart from this small effect, the profile of the polymer stripes is nearly rectangular, with an aspect ratio of 0.83. The high quality of the replication extends over the entire $100 \times 100 \mu\text{m}^2$ area that was covered by the master pattern.

conventional imprinting techniques². Several features of this electrostatic replication process should be pointed out. The quality of the replicated structures is nearly perfect and, to within the resolution of the AFM, the features have perpendicular side walls. The electrostatically induced structure formation in the polymer film is not limited to wavelength around λ_m ; rather, a large range in lateral structures can be replicated for otherwise identical experimental parameters. In the example shown here (Fig. 4), lines with a width of 140 nm were reproduced for an average plate spacing of 125 nm and an applied voltage of 42 V: the aspect ratio (height/width) of these structures is ~ 0.83 . The extension to lateral length scales of less than 100 nm and aspect ratios greater than 1 should be possible. We have shown that strong electric field can cause the breakup of dielectric fields, leading to characteristic hexagonal patterns. In combination with a patterned master, this principle can be employed to replicate submicrometre structures. This technique is straightforward, and does not need specialized equipment or specialized chemicals. In a technological application, the spacing between the two electrodes would need to be controlled with great precision, but this could be easily achieved by introducing non-conducting spacers into the capacitor gap. □

Received 22 September; accepted 9 December 1999.

1. *The National Technology Roadmap for Semiconductors* (Semiconductor Industry Association Report, 1997).
2. Xia, Y., Rogers, J. A., Paul, K. E. & Whitesides, G. M. Unconventional methods for fabricating and patterning nanostructures. *Chem. Rev.* **99**, 1823–1848 (1999).
3. Landau, L. D., Lifshitz, E. M. & Pitaevski, L. P. *Electrodynamics of Continuous Media* (Pergamon, Oxford, 1984).
4. Amundson, K. *et al.* Effect of an electric field on block copolymer microstructure. *Macromolecules* **24**, 6546–6548 (1991).
5. Amundson, K., Helfand, E., Quan, X. & Smith, S. D. Alignment of lamellar block copolymer microstructure in an electric field. 1. Alignment kinetics. *Macromolecules* **26**, 2698–2703 (1993).
6. Amundson, K., Helfand, E., Quan, X., Hudson, S. D. & Smith, S. D. Alignment of lamellar block copolymer microstructure in an electric field. 2. Mechanisms of alignment. *Macromolecules* **27**, 6559–6570 (1994).
7. Morkved, T. L. *et al.* Local control of microdomain orientation in diblock copolymer thin films with electric fields. *Science* **273**, 931–933 (1996).
8. Thurn-Albrecht, T., DeRouchey, J., Russell, T. P. & Jaeger, H. M. Overcoming interfacial interactions with electric fields. *Macromolecules* (submitted).
9. Blythe, A. R. *Electrical Properties of Polymers* (Cambridge Univ. Press, 1979).
10. Chou, S. Y., Zhuang, L. & Guo, L. Lithographically induced self-construction of polymer microstructures for resistless patterning. *Appl. Phys. Lett.* **75**, 1004–1006 (1999).
11. Chou, S. Y. & Zhuang, L. Lithographically induced self-assembly of periodic polymer micropillar arrays. *J. Vac. Sci. Technol. B* **17**, 3197–3202 (1999).
12. Reiter, G. Dewetting of thin polymer films. *Phys. Rev. Lett.* **68**, 75–78 (1992).
13. de Gennes, P. G. Wetting: statics and dynamics. *Rev. Mod. Phys.* **57**, 827–862 (1985).
14. Herminghaus, S. Dynamical instability of thin liquid films between conducting media. *Phys. Rev. Lett.* **83**, 2359–2361 (1999).

Acknowledgements

We thank J. Mlynek for support, S. Walheim for discussions, W. Zulehner and Wacker-Chemie GmbH for the silicon wafers, B. Maile and eXtreme Lithography for the master patterns, and C. Hawker for the brominated polystyrene. This work was supported by the Deutsche Forschungsgemeinschaft (DFG) through the Sonderforschungsbereich 513, NATO, the US Department of Energy and the National Science Foundation through the Materials Research Science and Engineering Center. T.T.A. acknowledges a research fellowship from the DFG.

Correspondence and requests for materials should be addressed to U.S. (e-mail: u.steiner@chem.rug.nl) or T.P.R. (e-mail: russell@mail.pse.umass.edu).

Variations of Younger Dryas atmospheric radiocarbon explicable without ocean circulation changes

Tomasz Goslar*, Maurice Arnold†, Nadine Tisnerat-Laborde‡, Justyna Czernik* & Kazimierz Włódkowski‡

* Institute of Physics, Silesian Technical University, 44-100 Gliwice, Poland

† Laboratoire des Sciences du Climat et de l'Environnement, CNRS-CEA, 91198 Gif sur Yvette, France

‡ Institute of Geography and Spatial Organization, Polish Academy of Sciences, 00-325 Warsaw, Poland

The concentration of radiocarbon, ^{14}C , in the atmosphere depends on its production rate by cosmic rays, and on the intensity of carbon exchange between the atmosphere and other reservoirs, for example the deep oceans. For the Holocene (the past ~11,500 years), it has been shown that fluctuations in atmospheric radiocarbon concentrations have been caused mostly by variations in the solar magnetic field^{1–3}. Recent progress in extending the radiocarbon record backwards in time^{4–10} has indicated especially high atmospheric radiocarbon concentrations in the Younger Dryas cold period, between 12,700 and 11,500 years before the present. These high concentrations have been interpreted as a result of a reduced exchange with the deep-ocean reservoir, caused by a drastic weakening of the deep-ocean ventilation^{7–9,11,12}. Here

we present a high-resolution reconstruction of atmospheric radiocarbon concentrations, derived from annually laminated sediments of two Polish lakes, Lake Gościąg and Lake Perespilno. These records indicate that the maximum in atmospheric radiocarbon concentrations in the early Younger Dryas was smaller than previously believed, and might have been caused by variations in solar activity. If so, there is no indication that the deep-ocean ventilation in the Younger Dryas was significantly different from today's.

The basic set of data consists of ^{14}C dates of the varved sediments of Lake Gościąg (LG). The LG dates published so far⁷ covered densely the time span between 11,800–10,500 calendar years before present, but only a few dates from before 11,800 cal. yr BP were available. In this work, we sampled the oldest part of the LG sediments, that is, between 12,950 and 11,700 cal. yr BP.

An additional data set comes from Lake Perespilno (LP), in eastern Poland. The sediments of that lake¹³ contain 3,100 varves in two sections. The younger one, covering the Younger Dryas, has been synchronized with the LG timescale, using the ^{14}C dates and pollen analysis¹³. Here, we completed ^{14}C measurements of the older section, and dated it to 14,455–13,645 (± 50) cal. yr BP, by matching the LP ^{14}C dates to the coral ^{14}C calibration data^{4–6}. The calendar and ^{14}C ages from Lakes Gościąg and Perespilno are provided in the Supplementary Information.

The LG dates (Fig. 1) perfectly match the high-precision ^{14}C calibration curve of German pines (GP; ref. 14), except for the oldest end of that curve. This would imply more than 200 varves missing from the LG chronology before 11.6 kyr BP, which is highly improbable as the chronology is replicated in many cores from three separate lake basins. Alternatively, the discrepancy might reflect problems with cross-correlation of the pine tree-ring sequences in that period, when few trunks are available, the tree-rings are very thin, and are frequently even missing¹⁵.

Except for a few outliers, the LG dates agree well with those obtained on corals^{4–6,16} and on the sediments of Lake Suigetsu (LS), Japan¹⁰. However, ^{14}C dates from the Cariaco Basin, Atlantic⁹, in the period between 13 and 12.5 kyr BP are significantly younger. This seems to be connected with the large difference between absolute ages of the Alleröd/Younger Dryas boundary in both archives (Fig. 2 and inset to Fig. 1). In all probability, one of two chronologies (either Cariaco or LG) has some error around the mid-Younger Dryas.

The ages of both Alleröd/Younger Dryas and Younger Dryas/Holocene boundaries in LG (ref. 8) perfectly agree with those estimated in the GRIP ice core^{17,18}. On the other hand, the chronology of the Cariaco sediments seems to be supported by that of the GISP2 ice core¹⁹. Direct comparison of proxy climatic records shows distinct time lags between several events (Fig. 2) recorded in both Greenland ice cores. This implies that either the GRIP or GISP2 chronology needs adjustment.

There are arguments for the correctness of the GRIP and LG time scales. The duration of the Younger Dryas in LG agrees very well with that in LP (ref. 13), and in the revised chronologies of German Maar lakes²⁰. ^{14}C dates of the oldest samples from LG (around 12.9 kyr BP) fitted most of the dates from corals and from the Lake Suigetsu, whereas in the same period, all the Cariaco dates are younger. Additionally, they offset the coral and LS dates in the whole preceding millennium. These arguments suggest that the Cariaco (and GISP2) timescales may not be adequate. Indeed, the ^{10}Be record from GISP2 (ref. 21) between 8 and 5 kyr BP, lags the dendro-dated record of atmospheric ^{14}C concentration by about 100 years, which is difficult to explain unless there is an error in GISP2 chronology³. In fact, the Younger Dryas/Holocene transitions in Cariaco and GISP2 are about 100 years older than in the GRIP, LG and GP archives.

The LG and LP data points (Fig. 1) have been connected by a spline curve. We realised that combining data from archives that are

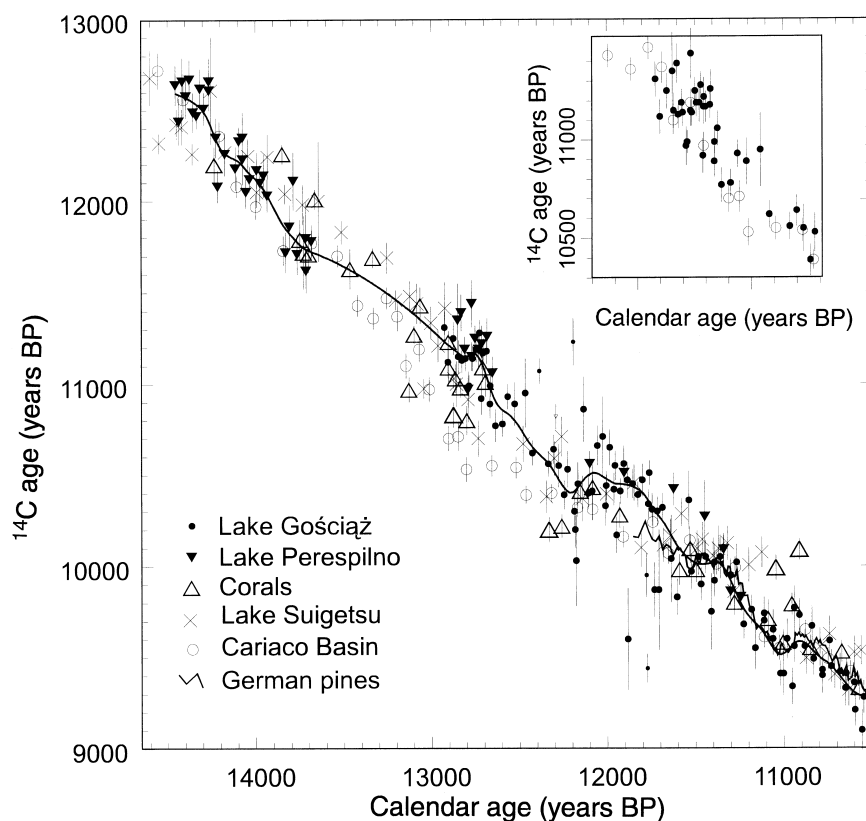


Figure 1 Radiocarbon versus calendar ages during the Late Glacial and early Holocene. Solid circles, sediments of Lake Gościąż (LG); inverted solid triangles, sediments of Lake Perespilno (LP); open triangles, corals from Barbados, Mururoa Atoll and Tahiti^{4,5,16}, and New Guinea⁶; crosses, sediments of Lake Suigetsu, Japan¹⁰; open circles, sediments from the Cariaco Basin, Atlantic Ocean⁹; solid line, German pines¹⁴. The LG and LP data published so far^{7,13} have been completed with 80 accelerator mass spectrometer (AMS) ¹⁴C dates, mainly from the oldest sections of sediments (see Supplementary Information). The absolute age of LG chronology⁸ was adjusted by 15 years, to synchronize the onset of climatic warming at the Younger Dryas/Preboreal transition, recorded in the LG sediments and in German pines¹⁵ (see Fig. 2). The smooth line is a cubic spline fitted to the Gościąż and Perespilno data points. The ¹⁴C dates that are outlying by more than 3 σ , not used for

the fit, have been marked by small symbols. Some dates are obviously too old, occurring in the earlier half of the Younger Dryas, when the vegetation cover around LG was the thinnest²⁶ and when the sediments contain some material rebedded from the lake littoral zone. Similarly rebedded sample is one from LP. Few dates (for samples less than 1 mg) seem to be contaminated with modern carbon. For the remaining 149 points, distribution of deviations of ¹⁴C dates from the spline curve is gaussian, and the mean-square deviation is 1.45 times the standard error of the ¹⁴C date, a reasonable value as the uncertainty of the calendar age of the samples is significant. Inset, good correspondence of the ¹⁴C dates from LG, LP (solid circles), and the Cariaco Basin (open circles), when the Cariaco basin Allerød/Younger Dryas transition (see Fig. 2) is synchronized with that recorded in the LG sediments.

not properly synchronized may produce artificial 'wiggles' in the curve; thus, we focused our attention on our data only. The curve features distinct plateaux in ¹⁴C age. The longest one, at 10,400 ¹⁴C BP, has not been reconstructed previously in detail. Another feature is the fast decline of ¹⁴C age at the onset of the Younger Dryas (12.7–12.5 kyr BP), which was found in the earlier studies^{7–9}. Our new dates make this decline much slower and more reasonable, as the high rate of decline required an extremely fast increase of atmospheric ¹⁴C concentration that is difficult to explain by known mechanisms (for example, ref. 8).

The wiggles of the ¹⁴C calibration curve reflect variations of atmospheric ¹⁴C concentration. Our record of $\Delta^{14}\text{C}$ (Fig. 3) differs from the previous reconstructions in two ways. First, oscillations of $\Delta^{14}\text{C}$ occur between 14.5 and 13.8 kyr BP. Second, the early-Younger Dryas maximum of $\Delta^{14}\text{C}$ seems distinctly lower than suggested previously^{7–9}, and it is followed by a rapid decline in $\Delta^{14}\text{C}$ around the middle of the Younger Dryas, an effect not documented in previous works.

The large maximum of atmospheric ¹⁴C concentration in the Younger Dryas has been interpreted as a result of drastic weakening of ventilation of deep oceans^{7–9,11,12}. However, most model studies had difficulties with simultaneous reproduction of the very high $\Delta^{14}\text{C}$ in the early Younger Dryas, and fast decline of $\Delta^{14}\text{C}$ in the second part of Younger Dryas. These two features were partly reconciled in the hypothesis of drastic weakening of the North

Atlantic Deep Water flux at the onset of the Younger Dryas, and strengthening of the North Atlantic Intermediate Water flux about 300 years later⁸. Nevertheless, the simulated $\Delta^{14}\text{C}$ curve still passed below most data points in the early Younger Dryas (see Fig. 6 in ref. 8), and above the points in the late Younger Dryas.

Our new data show (Fig. 3) that the early-Younger Dryas maximum of $\Delta^{14}\text{C}$ was smaller than previously believed. Such a maximum, and also the decline of $\Delta^{14}\text{C}$ between 11.6 and 11.3 kyr BP seem well explained by the scenario already proposed⁸. The most serious deviations of atmospheric $\Delta^{14}\text{C}$ from the model predictions are the minimum of $\Delta^{14}\text{C}$ between 12.1 and 11.7 kyr BP, and the maximum between 11.3 and 11.0 kyr BP. The minimum, if produced solely by the North Atlantic, would require a return to the normal flux of North Atlantic Deep Water, an event that is not likely to have occurred in the middle of the Younger Dryas. Another proposal is the strengthening of vertical circulation in the Southern Ocean²², raised to explain the climatic warming in Antarctica just when the Northern Hemisphere cooled down²³. However, the warming of Antarctica was gradual throughout the whole Younger Dryas, whereas the minimum of $\Delta^{14}\text{C}$ is short and occurs only in the middle of the Younger Dryas. The oceanic mechanism has also been raised to explain the increase of $\Delta^{14}\text{C}$ just before 11 kyr BP¹¹, during the so-called Preboreal Oscillation (PBO). However, this maximum is almost as large as that in the early Younger Dryas, whereas the climatic cooling in the PBO was much weaker than in the Younger

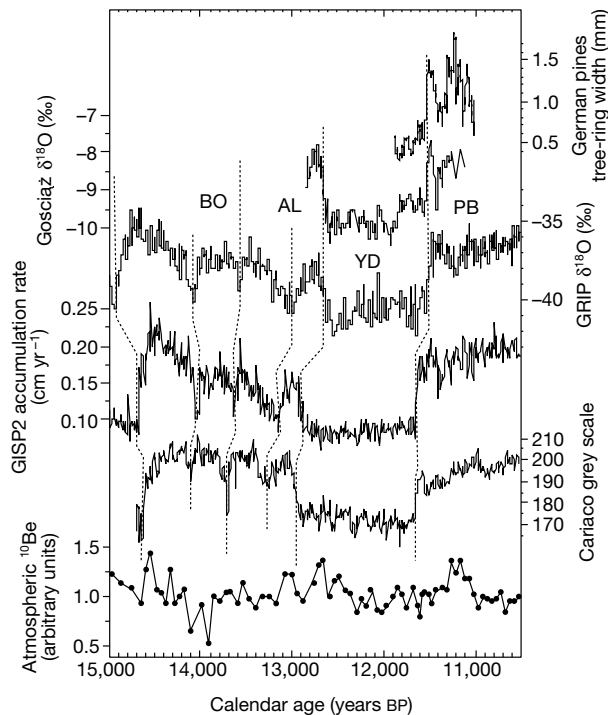


Figure 2 Comparison of calendar timescales of the records discussed in the text. The climatic events during the Late Glacial, common to all the records (such as the beginning of the Bölling/Alleröd transition, and the boundaries of the Younger Dryas) have been marked with dotted lines. BO, Bölling; AL, Alleröd; YD, Younger Dryas; PB, Preboreal.

Dryas. It is thus probable that some other factor (such as solar activity), or combination of factors, is responsible for the $\Delta^{14}\text{C}$ maxima in the Younger Dryas and Preboreal.

A suitable indicator of past changes of solar activity is the atmospheric concentration of ^{10}Be . It has been shown that the fluctuations of cosmogenic production rate, derived from the changes of ^{10}Be concentration in Greenland and Antarctic ice, explain medium-term $\Delta^{14}\text{C}$ variations in the Holocene^{23,21}. The reliability of a similar study in the Pleistocene²¹ is more problematic, mainly because of the varying snow accumulation rate which disturbs the direct relationship between ^{10}Be concentration in the ice and atmosphere. The flux of ^{10}Be to the ice sheet in the Greenland Summit has been shown²¹ to be roughly proportional to the snow accumulation rate, using the deviations from the linear relation to derive the changes of ^{10}Be concentration in the atmosphere (using the method of ref. 24). The record of ref. 21 hence shows the maxima in the early Younger Dryas and at about 11 kyr BP (Fig. 2). We used these changes to calculate the variations of $\Delta^{14}\text{C}$ in the early Holocene. The results (Fig. 3) strongly indicate that most of the Preboreal $\Delta^{14}\text{C}$ maximum may have been caused by an increase of cosmogenic production; thus, it is mostly of solar origin.

The ^{10}Be peak in the early Younger Dryas (Fig. 2) is quite similar to the Preboreal one, which suggests that the early-Younger Dryas $\Delta^{14}\text{C}$ maximum was also significantly influenced by fluctuation of solar activity. However, in the Late Glacial, both the snow accumulation rate, and the ^{10}Be concentration in the GISP2 ice, vary greatly. As a result, the accuracy of atmospheric ^{10}Be reconstruction in that period is low and difficult to assess. An additional complication is caused by the apparent stretching of the GISP2 timescale within the Younger Dryas, which would strongly affect the derived levels of atmospheric ^{10}Be . Therefore, using ^{10}Be to explain the mechanisms of $\Delta^{14}\text{C}$ variations during the Late Glacial seems premature.

Simulations of the effect of melt water on the thermohaline circulation, performed with general circulation models, suggest an increase of atmospheric $\Delta^{14}\text{C}$ by about 30‰ (ref. 12) or by only 10–15‰ (ref. 25); the latter value was obtained when the

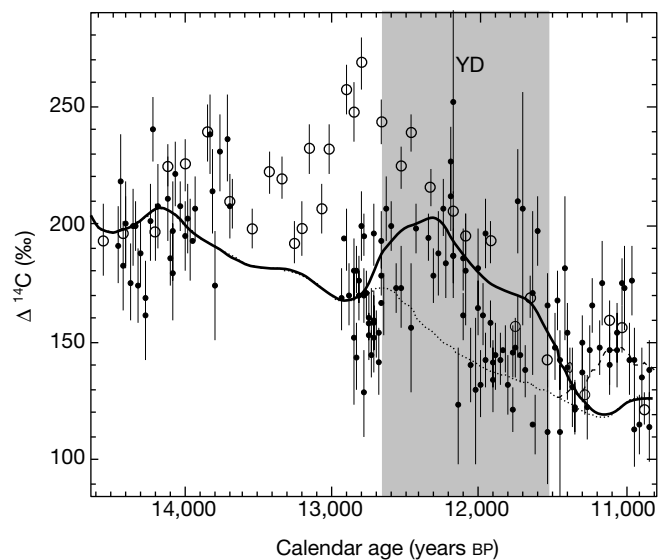


Figure 3 Atmospheric ^{14}C concentrations in the Late Glacial and early Holocene. Solid symbols, $\Delta^{14}\text{C}$ data from Lake Gosclaz and Peresipino; open circles, $\Delta^{14}\text{C}$ data from the Cariaco basin⁹. The smooth solid line represents variations of $\Delta^{14}\text{C}$ produced by the changes of geomagnetic field, CO_2 content in the atmosphere and the vertical oceanic circulation, calculated in ref. 8, with a slightly adjusted mean value of ^{14}C production rate. In the standard model the ^{14}C production rate which corresponds to the current geomagnetic field maintains the steady-state $\Delta^{14}\text{C} = 0$ ‰. Here, we used the production rate lowered by 5% (ref. 7). The dotted line shows the trend of $\Delta^{14}\text{C}$, produced solely by changes in the Earth's magnetic field and CO_2 content in the atmosphere. The dashed line represents variations of atmospheric $\Delta^{14}\text{C}$, explained by fluctuations of ^{14}C production rate in the early Holocene. These variations depend on the production of ^{14}C (controlled by the solar as well as geomagnetic fields), integrated over the whole globe. The ^{10}Be concentration in polar ice, on the other hand, is proportional to cosmic-ray intensity at high latitudes. In Antarctica, geomagnetic modulation of cosmic rays is negligible, and the effects of solar modulation are about 1.4 times stronger than for the whole atmosphere³. In our calculations (performed with the PANDORA model²⁷) we assumed the variations of ^{14}C production rate to be proportional to fluctuations of atmospheric ^{10}Be concentration²¹ around the mean value (divided by a factor of 1.4), multiplied by the geomagnetically controlled changes of cosmogenic production (the latter ones used earlier in refs 7–9). The air over Greenland is much more influenced by the aerosols from low latitudes than on the Antarctic plateau. For that reason the local gain of solar modulation in the GISP2 ^{10}Be record may be lower than 1.4, and the actual variations of $\Delta^{14}\text{C}$ even stronger than calculated here.

weakening of the North Atlantic circulation was counterbalanced by the increasing ventilation of the deep Southern Ocean. If this was in fact the case, it seems that changes of oceanic circulation during the Younger Dryas are beyond our detection limit, until independent precise data on the ^{14}C production rate are available. □

Received 23 April; accepted 21 December 1999.

1. Stuiver, M. & Quay, P. D. Changes in atmospheric carbon-14 attributed to a variable Sun. *Science* **207**, 11–19 (1980).
2. Beer, J. *et al.* Information on past solar activity and geomagnetism from ^{10}Be in the Camp Century ice core. *Nature* **331**, 675–679 (1988).
3. Bard, E., Raisbeck, G. M., Yiou, F. & Jouzel, J. Solar modulation of cosmogenic nuclide production over the last millennium: comparison between ^{14}C and ^{10}Be records. *Earth Planet. Sci. Lett.* **150**, 453–462 (1997).
4. Bard, E. Geochemical and geophysical implications of the radiocarbon calibration. *Geochim. Cosmochim. Acta* **62**, 2025–2038 (1998).
5. Bard, E. *et al.* Deglacial sea-level record from Tahiti corals and the timing of global meltwater discharge. *Nature* **382**, 241–244 (1996).
6. Edwards, R. L. *et al.* A large drop in atmospheric $^{14}\text{C}/^{12}\text{C}$ and reduced melting in the Younger Dryas, documented with ^{230}Th ages of corals. *Science* **260**, 962–968 (1993).
7. Goslar, T. *et al.* High concentration of atmospheric ^{14}C during the Younger Dryas cold episode. *Nature* **377**, 414–417 (1995).
8. Goslar, T. *et al.* Variations of atmospheric ^{14}C concentrations over the Alleröd-Younger Dryas transition. *Clim. Dyn.* **15**, 29–42 (1999).
9. Hughen, K. A. *et al.* Deglacial changes in ocean circulation from an extended radiocarbon calibration. *Nature* **391**, 65–68 (1998).

10. Kitagawa, H. & van der Plicht, J. Atmospheric radiocarbon calibration to 45,000 yr B.P.: late glacial fluctuations and cosmogenic isotope production. *Science* **279**, 1187–1190 (1998).
11. Björck, S. *et al.* Synchronised terrestrial-atmospheric deglacial records around the North Atlantic. *Science* **274**, 1155–1160 (1996).
12. Stocker, T. F. & Wright, D. G. Rapid changes in ocean circulation and atmospheric radiocarbon. *Paleoceanography* **11**, 773–795 (1996).
13. Goslar, T. *et al.* Climate-related variations in the composition of the late glacial and early Holocene sediments of Lake Perespinlo (eastern Poland). *Quat. Sci. Rev.* **18**, 899–911 (1999).
14. Stuiver, M. *et al.* INTCAL98 Radiocarbon age calibration, 24,000–0 cal BP. *Radiocarbon* **40**, 1041–1083 (1998).
15. Spurk, M. *et al.* Revisions and extension of the Hohenheim oak and pine chronologies: new evidence about the timing of the Younger Dryas/Preboreal transition. *Radiocarbon* **40**, 1107–1116 (1998).
16. Bard, E., Hamelin, B., Fairbanks, R. G. & Zindler, A. Calibration of the ^{14}C timescale over the past 30,000 years using mass spectrometric U–Th ages from Barbados corals. *Nature* **345**, 405–410 (1990).
17. Johnsen, S. J. *et al.* Irregular glacial interstadials recorded in a new Greenland ice core. *Nature* **359**, 311–313 (1992).
18. Hammer, C. U. *et al.* *The Stratigraphic Dating of the GRIP Ice Core*. (Special Report of the Geophysics Department, Inst. Niels Bohr for Astronomy, Physics and Geophysics, Univ. Copenhagen, 1998).
19. Alley, R. B. Abrupt increase in Greenland snow accumulation at the end of the Younger Dryas event. *Nature* **362**, 527–529 (1993).
20. Brauer, A. *et al.* High resolution sediment and vegetation responses to Younger Dryas climate change in varved lake sediments from Meerfelder Maar, Germany. *Quat. Sci. Rev.* **18**, 321–329 (1999).
21. Finkel, R. C. & Nishiizumi, K. Beryllium 10 concentrations in the Greenland Ice Sheet Project 2 ice core from 3–40 ka. *J. Geophys. Res.* **102**, 26699–26706 (1997).
22. Broecker, W. S. Paleocene circulation during the last deglaciation: A bipolar seesaw? *Paleoceanography* **13**, 119–121 (1998).
23. Blunier, T. *et al.* Timing of the Atlantic cold reversal and the atmospheric CO_2 increase with respect to the Younger Dryas event. *Geophys. Res. Lett.* **24**, 2683–2686 (1997).
24. Alley, R. B. *et al.* Changes in continental and sea-salt atmospheric loadings in central Greenland during the most recent deglaciation: Model-based estimates. *J. Glaciology* **41**, 503–514 (1995).
25. Mikolajewicz, U. A Meltwater Induced Collapse of the “Conveyor Belt” Thermohaline Circulation and its Influence on the Distribution of $\Delta^{14}\text{C}$ and $\delta^{18}\text{O}$ in the Oceans (Technical Report 189, Max-Planck-Institut für Meteorologie, Hamburg, 1996).
26. Goslar, T. *et al.* High-resolution lacustrine record of the Late Glacial/Holocene transition in Central Europe. *Quat. Sci. Rev.* **12**, 287–294 (1993).
27. Broecker, W. S. *et al.* The distribution of radiocarbon in the Glacial Ocean. *Glob. Biogeochem. Cycles* **4**, 103–117 (1990).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank R. Finkel for the numerical data on ^{10}Be concentration in the GISP2 ice core, R. Alley for the data on the GISP2 accumulation rate, K. Hughen for the ^{14}C calibration data and the grey-scale record of the Cariaco sediments, S. Johnsen and C. Hammer for the data on the GRIP timescale, M. Spurk for the data on the German pine tree-ring thickness, and E. Bard for comments on the manuscript. The sediment cores from Lake Gościąg and Lake Perspino were raised with the help of J. Pawlyta, A. Rakowski and A. Walanus. The terrestrial macrofossils were picked up by D. Moszyńska-Moskwa. The French Centre National de la Recherche Scientifique et Commissariat de l'Energie Atomique carried out accelerator mass spectrometer (AMS) ^{14}C measurements. This work has been sponsored by the Polish Committee for Scientific Research, through the grant to T.G.

Correspondence and requests for materials should be addressed to T.G. (e-mail: goslar@zeus.polsl.gliwice.pl).

A thermodynamic explanation for black smoker temperatures

Tim Jupp & Adam Schultz

Institute of Theoretical Geophysics, Department of Earth Sciences, University of Cambridge, Cambridge CB2 3EQ, UK

There is a remarkable difference between the maximum temperature of black smoker effluent (350°C – 400°C) and the temperature of the solidifying magma which heats it ($\sim 1,200^\circ\text{C}$)^{1–3}. It has been suspected⁴ for some time that the nonlinear thermodynamic properties of water⁵ might be responsible for this discrepancy. Here, we translate this hypothesis into a physical model, by examining the internal temperature structure of convection cells in a porous medium. We demonstrate that, at pressures appropriate to seafloor crust, plumes of pure water form naturally at

$\sim 400^\circ\text{C}$ for any heat source with temperature greater than $\sim 500^\circ\text{C}$. Higher temperatures are confined to a boundary layer at the base of the convection cell, where the flow is horizontal. The phenomenon is explained analytically using the thermodynamic properties of water, and is illustrated by numerical simulations. Our model predicts the existence of the high-temperature ‘reaction zone’ found in ophiolites⁶ and suggests that vent temperatures will remain steady as magma chambers solidify and cool⁷.

Three mechanisms to limit black smoker temperatures have been proposed. The first suggestion is that high-temperature rock (greater than, say, 500°C) is effectively impermeable, because its ductile nature prevents the formation of cracks^{8,9}. The second is that high-temperature rock becomes impermeable over time, either by mineral precipitation^{4,10}, or by thermal expansion¹¹. The third hypothesis is that the temperature cap is imposed by the thermodynamic properties of water. The first two mechanisms work by restricting flow to sufficiently cool regions of the crust. In contrast, the third mechanism allows fluid to flow at all temperatures within the crust, but limits the temperature of the fluid expelled at the sea floor.

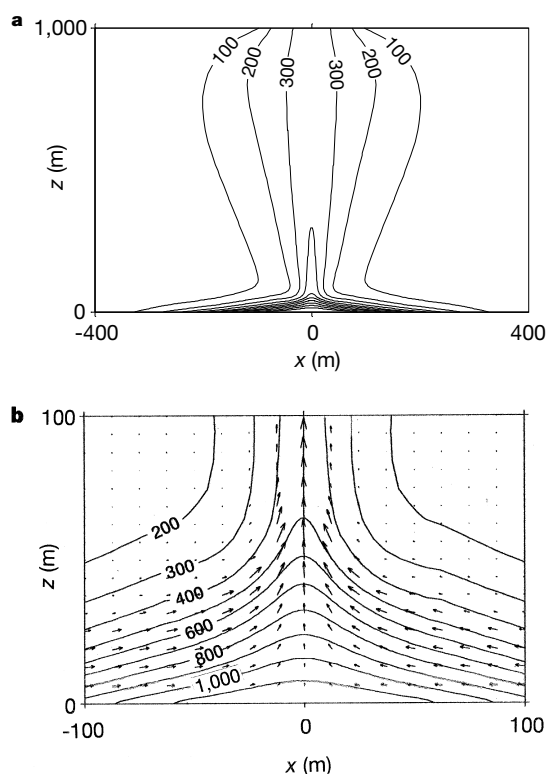


Figure 1 The steady-state temperature distribution in a convection cell at sea floor pressures. The top boundary ($z = 1,000\text{ m}$) represents the seafloor and is permeable. Here, the pressure is held at 25 MPa (equivalent to $\sim 2.5\text{ km}$ below the sea surface). The temperature boundary condition is $T = 10^\circ\text{C}$ where flow is downwards, and $\partial T/\partial z = 0$ where flow is upwards. The side boundaries for the simulation are at $x = -1,700\text{ m}$ and $x = 1,700\text{ m}$ (beyond the range of these figures), sufficiently distant that they do not influence the solution near the heater. They are held at $T = 10^\circ\text{C}$ and cold hydrostatic pressure. The bottom boundary ($z = 0$) is impermeable. A gaussian (bell-shaped) temperature profile is imposed, representing a magma chamber. The temperature on $z = 0$ runs from ‘cold’ (10°C) at the sides to ‘hot’ ($1,200^\circ\text{C}$) in the centre. Parameters for the simulation are a porosity of 10%, permeability $k = 10^{-14}\text{ m}^2$, thermal conductivity $\lambda = 2.0\text{ W m}^{-1}\text{ K}^{-1}$. **a**, The overall temperature structure of the convection cell, showing the distinction between the boundary layer and the plume. Isotherms from 100°C to $1,100^\circ\text{C}$ are drawn, in increments of 100°C . Flow vectors are omitted for clarity. Flow is downwards at the sides, towards the centre at the base and upwards in the centre. **b**, A close-up view of the flow regime and temperature structure inside the boundary layer, showing the bottom 100 m of the domain. Isotherms from 200°C to $1,100^\circ\text{C}$ are drawn, in increments of 100°C . Vectors are of Darcy velocity.

Previous studies have shown how the thermodynamic properties of water affect the onset of convection¹² and the overall rate of heat transfer^{13,14} in convection cells operating across a small temperature difference ($\sim 10^\circ\text{C}$). Given a fixed temperature difference, convection cells operating near the critical point of water ($\sim 22\text{ MPa}$, $\sim 374^\circ\text{C}$) transfer heat much more rapidly than cells at other temperatures. Such 'superconvection' suggests that hydrothermal systems should be able to transfer heat effectively, but it is insufficient to explain the limit to vent temperatures. Real systems operate between $\sim 10^\circ\text{C}$ and $\sim 1,200^\circ\text{C}$ —a large temperature difference—and vent at a particular temperature ($\sim 400^\circ\text{C}$). To explain this, we investigate the internal temperature structure of the cell rather than the overall heat transfer.

A numerical simulation can be used to illustrate the thermal structure of convection cells¹⁵ (Fig. 1). This simulation uses parameter values appropriate to seafloor systems. However, analysis of the governing equations shows that the main feature of the temperature structure depends solely on the thermodynamics of water, and not on the details of the parameterization. To isolate the phenomenon under discussion, all previously suggested temperature-limiting mechanisms are excluded, and the crust is taken to be homogeneous and isotropic. Simulations were performed with the software package HYDROTHERM¹⁶ (available at <http://water.usgs.gov/software/hydrotherm.html>), modified to enforce a thermal boundary condition that is relevant to seafloor systems¹⁷. (Hot effluent emerges from an otherwise cold seafloor, so a fixed-temperature boundary condition is not always appropriate. Consequently, the condition $\partial T/\partial z = 0$ is imposed on the sea floor where flow is upwards, and $T = 10^\circ\text{C}$ is imposed where flow is downwards).

The governing equations are Darcy's law¹⁸, an equation of state for pure water^{19–22} and the conservation of mass and energy in a porous medium¹⁸. Convection is initiated by a 'heater' along part of the bottom boundary²³, which represents the top of a magma chamber. For simplicity, the heater imposes a gaussian (bell-shaped) temperature profile which runs from 'cold' (T_0) at the sides to 'hot' ($T_0 + \Delta T$) at the centre. The heater is 'switched on' at time zero, when the domain is cold ($T = T_0$) and at hydrostatic pressure. Time-dependent solutions for the temperature and pressure fields are calculated and allowed to evolve until a steady state is reached.

A typical steady-state solution with $T_0 = 10^\circ\text{C}$ and $(T_0 + \Delta T) = 1,200^\circ\text{C}$ is shown in Fig. 1. There is a thin boundary layer of very hot fluid (500°C to $1,200^\circ\text{C}$) at the base of the convection cell. This layer, in which flow is horizontal, can be compared with the 'reaction zone' observed in ophiolites, for which similar tempera-

tures have been inferred⁶. The temperature of the plume is about 400°C . Similar plume temperatures were obtained for other systems dominated by convection, with $T_0 = 10^\circ\text{C}$ and $(T_0 + \Delta T) > \sim 500^\circ\text{C}$, including systems with anisotropic permeability. It follows that vent temperatures could remain steady as a magma chamber solidified and cooled, dropping only when the heat source fell below $\sim 500^\circ\text{C}$.

It appears that plumes have a natural tendency to form at $\sim 400^\circ\text{C}$, given a sufficient heat source and seafloor pressures. The explanation for this phenomenon lies in the balance between conduction and advection—the two mechanisms which transport energy in a convection cell.

In the model (Fig. 1), the temperature T and pressure p are functions of position $\mathbf{x} = (x, z)$. Thermal gradients produce a conductive heat flux $-\lambda \nabla T$, where λ is the thermal conductivity. Fluid motion creates an advective heat flux $\rho \mathbf{h} \mathbf{u}$, for water of density ρ , specific enthalpy h and Darcy velocity $\mathbf{u}$. In the conservation of energy equation¹⁸, the rate of change of energy per unit volume is given by the negative divergence of the total heat flux, or $-\nabla \cdot (\rho \mathbf{h} \mathbf{u} - \lambda \nabla T)$. This can be split into two terms, $-\nabla \cdot (\rho \mathbf{h} \mathbf{u})$ and $\lambda \nabla^2 T$, which are rates of accumulation of energy per unit volume, due to advection and conduction respectively. At any point in the cell, the relative importance of these two mechanisms can be expressed by the ratio $Ra_L(\mathbf{x}, t) = |\nabla \cdot (\rho \mathbf{h} \mathbf{u})| / \lambda \nabla^2 T$. This ratio measures the influence of fluid motion on the evolution of the local temperature field, and could logically be labelled the 'local Rayleigh number'. (Traditionally, a single parameter—the 'Rayleigh number'—is defined for the whole cell¹². Here, a 'local Rayleigh number' has been assigned, independently, to each point in the cell.) If $Ra_L \ll 1$, conduction dominates over advection, and the temperature evolves independently of any fluid flow. Conversely, where $Ra_L \gg 1$, advection is strong enough to influence the temperature field, and the isotherms can be distorted away from a pattern dominated by conduction. Advective distortion of the isotherms marks the separation of the plume from the boundary layer—indeed, it is what defines the plume on a plot of the temperature field (Fig. 1). To discover where plumes will appear, it is important to know where Ra_L is largest.

After the heater is switched on, a thermal signal propagates upwards (Fig. 2). For the period before plume formation (Fig. 2a and b), the approximate magnitudes of the energy accumulation terms can be calculated. Let d be the distance that the signal has penetrated into the domain, that is, the vertical range over which the temperature drops from 'hot' ($T_0 + \Delta T$) to 'cold' (T_0). Suppose that $|\rho \mathbf{h} \mathbf{u}|$ is of order Φ . Hence $|\nabla \cdot (\rho \mathbf{h} \mathbf{u})|$ is of order Φ/d , and $|\lambda \nabla^2 T|$ is of order $\lambda \Delta T/d^2$. Consequently, Ra_L is of order $(\Phi/\lambda \Delta T) \cdot d$, and so

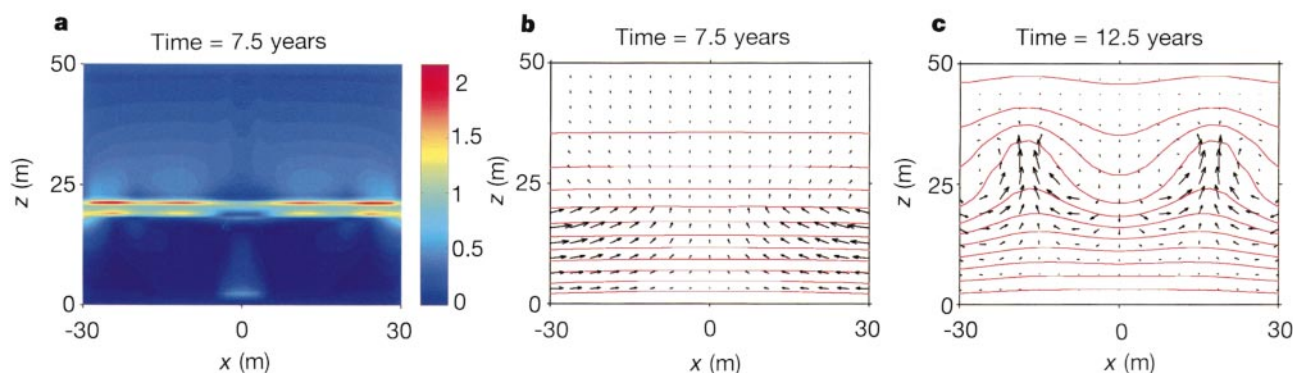


Figure 2 Early stages of the simulation, at the bottom of the domain shown in Fig. 1. **a**, Colour plot of the local Rayleigh number Ra_L , before plume formation. The regions of greatest Ra_L indicate where plumes are likely to form. **b**, The temperature and flow fields before plume formation. The thermal structure is governed by conduction alone. Convective motion exists, but it does not influence the temperature field. **c**, The

temperature and flow fields after plume formation. Advection has begun to distort the thermal structure. (As the simulation progresses to steady state, these two plumes coalesce to form the single plume in Fig. 1.) Isotherms are drawn from 100°C (top) to $1,100^\circ\text{C}$ (bottom) in increments of 100°C . Vectors are of Darcy velocity.

scales linearly with the vertical length scale d . Just before plume formation, d is sufficiently small that Ra_L is everywhere of order unity (Fig. 2a). Note that Ra_L is greatest near the 400 °C isotherm. A short time later, when d , and hence Ra_L , have increased slightly, plumes appear at about 400 °C (Fig. 2c).

By making some simplifications, it is possible to estimate the temperature at which Ra_L is maximized, just before plume formation. Temperature gradients are nearly uniform (Fig. 2b). It follows that the maximum value of Ra_L is controlled by the maximum value of $|\nabla \cdot (\rho \mathbf{h} \mathbf{u})|$. By analogy with convection in a U-shaped pipe²⁴, a convection cell can be divided into a cold 'downflow limb' and a hot 'upflow limb'. The vertical pressure gradient must be sufficiently small for cold water to flow down the downflow limb, and sufficiently large for hot water to flow up the upflow limb. Hence, the pressure gradient lies between hot hydrostatic and cold hydrostatic. The permeability is the same in both limbs, but the total resistance to flow is much greater in the upflow limb because of its

smaller cross-sectional area (Fig. 1). It follows that the pressure gradient will be much closer to cold hydrostatic than hot hydrostatic. Hence $\partial p / \partial z \approx -\rho_0 g$, where ρ_0 is the density of cold water. Suppose that the advective heat flux $\rho \mathbf{h} \mathbf{u}$ is dominated by its vertical component, which by Darcy's law¹⁸ is $-k(\partial p / \partial z + \rho g) \rho h / \mu$, for permeability k , gravitational acceleration g , and dynamic viscosity μ . An approximate expression for the advective energy accumulation is then:

$$-\nabla \cdot (\rho \mathbf{h} \mathbf{u}) \approx \left[\frac{\partial}{\partial z} \left(\frac{(\rho_0 - \rho) \rho h}{\mu} \right) \right] \cdot [-gk]$$

Denoting $(\rho_0 - \rho) \rho h / \mu$ by F , it follows that Ra_L should be maximized where $|\partial F / \partial z|$ is maximized. The quantity F has been termed 'fluxibility'¹ and measures the ability of buoyancy-driven water to transport energy. Fluxibility is a function of the thermodynamic state of the water (ρ_0 is constant and ρ , μ and h are known as functions of pressure and temperature from steam tables^{19–22}). Figure 3a shows how fluxibility depends on pressure and temperature over an appropriate range. In a seafloor hydrothermal system, the fluxibility varies greatly with temperature, but little with pressure. Before the formation of a plume, temperature is approximately a linear function of height (Fig. 2b). Thus, $|\partial F / \partial z|$ should be maximized roughly where $|\partial F / \partial T|$ is maximized.

Figure 3c shows clear peaks in $|\partial F / \partial T|$ at ~ 384 °C ($p = 25$ MPa) and ~ 412 °C ($p = 35$ MPa). These are the temperatures at which plumes would separate from the boundary layer under the simplifying assumptions. They are remarkably close to the plume temperatures in the numerical simulation (Fig. 1a and b, Fig. 2c).

This analysis uses the properties of pure water, because values of ρ , μ and h are not available for salt water over all of the required temperature range^{4,25}. When such data become available, the calculation should be repeated for salt water. It seems likely⁴ that the peaks in the $|\partial F / \partial T|$ curves would be shifted to higher temperatures by about 30 °C. Nonetheless, the use of pure water has one definite advantage. At pressures greater than 22 MPa (seafloor depth ~ 2.2 km), pure water is a single-phase fluid. Hence, the complication of phase separation has been removed from this analysis, and it has been possible to demonstrate a temperature-limiting mechanism that is independent of 'phase-separation' or 'two-phase effects'.

Three differences between this approach and classical studies of porous medium convection²⁶ should be noted. First, convection is not here generated by a uniformly heated lower boundary. The presence of both 'hot' and 'cold' at the base of the cell ensures that there is always a horizontal temperature gradient and associated fluid flow. The point of interest is not the onset of convective motion, but the onset of convective distortion of the temperature field. Second, no use is made of the Boussinesq approximation²⁷, under which the advective heat flux would be simplified to $\rho_0 \mathbf{h} \mathbf{u}$. The true value $\rho \mathbf{h} \mathbf{u}$ is used here because ρ can vary from ρ_0 by a factor of 25. Thirdly, this analysis abandons the traditional Rayleigh number²⁷, which is an overall balance between advection and conduction, in favour of a local Rayleigh number giving the balance at each point in the cell. This permits a discussion of the precise position (equivalently, temperature) at which plumes form.

At a given pressure, and given a sufficient source of heat, a plume is likely to form at the temperature of maximum fluxibility gradient (Fig. 3c). Hence, if ambient pressures are known, plume formation temperatures could be inferred for single-phase hydrothermal systems. Vents may be cooler than this plume-formation temperature because of conductive losses from the ascending fluid, but they cannot be hotter.

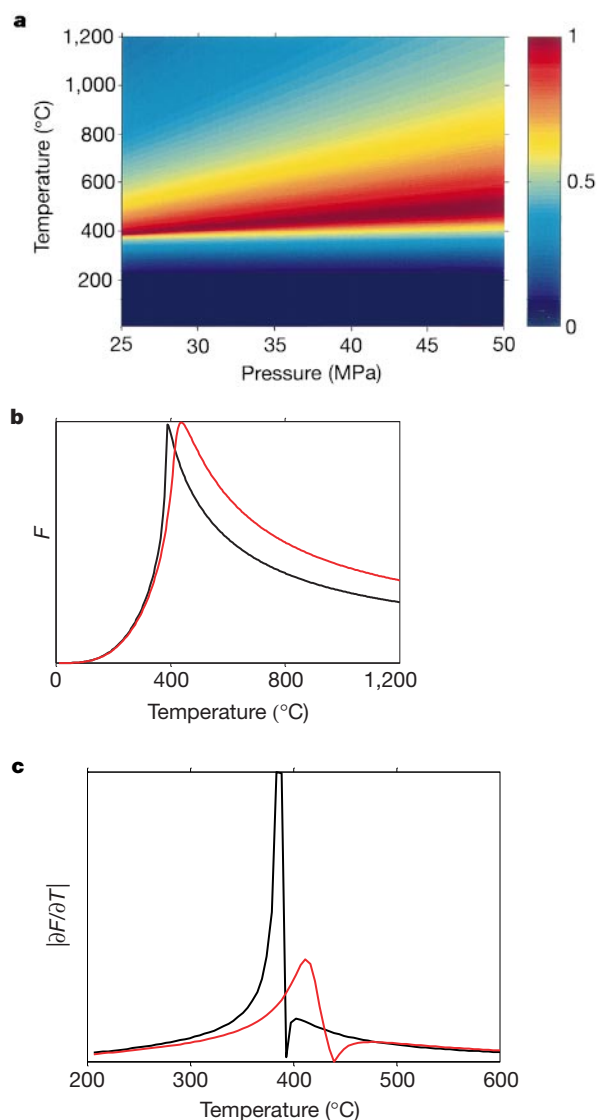


Figure 3 'Fluxibility' F and 'fluxibility gradient' $|\partial F / \partial T|$ (in normalized units) **a**, F as a function of temperature T and pressure p . The pressure range of 25 MPa to 50 MPa (~ 2.5 km to ~ 5 km cold hydrostatic head) is representative of the pressures found in seafloor hydrothermal systems. **b**, F as a function of T at $p = 25$ MPa (black line) and $p = 35$ MPa (red line). These pressures are lower and upper bounds for the pressures to be found in the cell in Fig. 1. **c**, $|\partial F / \partial T|$ as a function of T at $p = 25$ MPa (black line) and $p = 35$ MPa (red line). The temperature at which $|\partial F / \partial T|$ is maximized is, approximately, the temperature at which plumes form.

Received 6 May; accepted 9 December 1999.

1. Lister, C. R. B. Heat transfer between magmas and hydrothermal systems, or, six lemmas in search of a theorem. *Geophys. J. Int.* **120**, 45–59 (1995).
2. Lowell, R. P., Rona, P. A. & Von Herzen, R. P. Seafloor hydrothermal systems. *J. Geophys. Res.* **100**, 327–352 (1995).
3. Wilcock, W. S. D. Cellular convection models of mid-ocean ridge hydrothermal circulation and the

- temperatures of black smoker fluids. *J. Geophys. Res.* **103**, 2585–2596 (1998).
4. Bischoff, J. L. & Rosenbauer, R. J. An empirical equation of state for hydrothermal seawater (3.2 percent NaCl). *Amer. J. Sci.* **285**, 725–763 (1985).
 5. Johnson, J. W. & Norton, D. Critical phenomena in hydrothermal systems: state, thermodynamic, electrostatic, and transport properties of H₂O in the critical region. *Am. J. Sci.* **291**, 541–648 (1991).
 6. Gillis, K. M. & Roberts, M. D. Cracking at the magma-hydrothermal transition: evidence from the Troodos Ophiolite, Cyprus. *Earth Planet. Sci. Lett.* **169**, 227–244 (1999).
 7. Lowell, R. P. & Germanovich, L. N. On the temporal evolution of high-temperature hydrothermal systems at ocean ridge crests. *J. Geophys. Res.* **99**, 565–575 (1994).
 8. Lister, C. R. B. On the penetration of water into hot rock. *Geophys. J. R. Astron. Soc.* **39**, 465–509 (1974).
 9. Fournier, R. O. The transition from hydrostatic to greater than hydrostatic fluid pressure in presently active continental hydrothermal systems in crystalline rock. *Geophys. Res. Lett.* **18**, 955–958 (1991).
 10. Lowell, R. P., Van Cappellen, P. & Germanovich, L. Silica precipitation in fractures and the evolution of permeability in hydrothermal upflow zones. *Science* **260**, 192–194 (1993).
 11. Germanovich, L. N. & Lowell, R. P. Percolation theory, thermoelasticity, and discrete hydrothermal venting in the Earth's crust. *Science* **255**, 1564–1567 (1992).
 12. Straus, J. M. & Schubert, G. Thermal convection of water in a porous medium: effects of temperature- and pressure-dependent thermodynamic and transport properties. *J. Geophys. Res.* **82**, 325–333 (1977).
 13. Dunn, J. C. & Hardee, H. C. Superconvecting geothermal zones. *J. Volcanol. Geotherm. Res.* **11**, 189–201 (1981).
 14. Ingebritsen, S. E. & Hayba, D. O. Fluid flow and heat transport near the critical point of H₂O. *Geophys. Res. Lett.* **21**, 2199–2202 (1994).
 15. Norton, D. & Knight, J. Transport phenomena in hydrothermal systems: cooling plutons. *Am. J. Sci.* **277**, 937–981 (1977).
 16. Hayba, D. O. & Ingebritsen, S. E. *The Computer Model HYDROTHERM, a Three-dimensional Finite-difference Model to Simulate Ground-water Flow and Heat Transport in the Temperature Range of 0 to 1,200 degrees Celsius*. (US Geological Survey Water-Resources Investigations Report, 94-4045, Denver, Colorado, 1994).
 17. Dickson, P., Schultz, A. & Woods, A. in *Hydrothermal Vents and Processes* (ed. Parson, L. M. et al.) Vol. **18**, 145–157 (Geological Society Special Publication, London, 1995).
 18. Faust, C. R. & Mercer, J. W. Geothermal reservoir simulation 1. Mathematical models for liquid- and vapor-dominated hydrothermal systems. *Wat. Resour. Res.* **15**, 23–30 (1979).
 19. Haar, L., Gallagher, J. S. & Kell, G. S. *NBS/NRC Steam Tables*, 1–320 (Hemisphere, New York, 1984).
 20. Watson, J. T. R., Basu, R. S. & Sengers, J. V. An improved representative equation for the dynamic viscosity of water substance. *J. Phys. Chem. Ref. Data* **9**, 1255–1290 (1980).
 21. Sengers, J. V. & Kamgar-Parsi, B. Representative equations for the viscosity of water substance. *J. Phys. Chem. Ref. Data* **13**, 185–205 (1984).
 22. Sengers, J. V. & Watson, J. T. R. Improved international formulation for the viscosity and thermal conductivity of water substance. *J. Phys. Chem. Ref. Data* **15**, 1291–1314 (1986).
 23. Elder, J. E. *Geothermal Systems* 180–187 (Academic, London, 1981).
 24. Pascoe, A. R. & Cann, J. R. in *Hydrothermal Vents and Processes* (ed. Parson, L. M. et al.) Vol. **87**, 159–173 (Geological Society Special Publication, London, 1995).
 25. Anderko, A. & Pitzer, K. S. Equation-of-state representation of phase equilibria and volumetric properties of the system NaCl–H₂O above 573 K. *Geochim. Cosmochim. Acta* **57**, 1657–1680 (1993).
 26. Lapwood, E. R. Convection of a fluid in a porous medium. *Proc. Camb. Phil. Soc.* **44**, 508–521 (1948).
 27. Phillips, O. M. *Flow and Reactions in Permeable Rocks* 245–264 (Cambridge Univ. Press, 1991).

Acknowledgements

We thank USGS for permission to use HYDROTHERM. The calculations were carried out on the Enigma high-performance computing centre at the Institute of Theoretical Geophysics. We acknowledge the support of that facility by the Higher Education Research Foundation for England. T.J. is supported by a studentship from the Natural Environment Research Council.

Correspondence and requests for materials should be addressed to T.J. (e-mail: tim@itg.cam.ac.uk).

An interconnected network of core-forming melts produced by shear deformation

D. Bruhn, N. Groebner & D. L. Kohlstedt

Department of Geology & Geophysics, University of Minnesota, Minneapolis, Minnesota 55455, USA

The formation mechanism of terrestrial planetary cores is still poorly understood, and has been the subject of numerous experimental studies^{1–3}. Several mechanisms have been proposed by which metal—mainly iron with some nickel—could have been extracted from a silicate mantle to form the core. Most recent

models involve gravitational sinking of molten metal or metal sulphide through a partially or fully molten mantle^{4,5} that is often referred to as a 'magma ocean'. Alternative models invoke percolation of molten metal along an interconnected network (that is, porous flow) through a solid silicate matrix^{6,7}. But experimental studies performed at high pressures^{1–3} have shown that, under hydrostatic conditions, these melts do not form an interconnected network, leading to the widespread assumption that formation of metallic cores requires a magma ocean. In contrast, here we present experiments which demonstrate that shear deformation to large strains can interconnect a significant fraction of initially isolated pockets of metal and metal sulphide melts in a solid matrix of polycrystalline olivine. Therefore, in a dynamic (non-hydrostatic) environment, percolation remains a viable mechanism for the segregation and migration of core-forming melts in a solid silicate mantle.

Percolation of core-forming melts by porous flow through a silicate matrix has until now been largely ruled out as a possible melt segregation mechanism: this has been on the basis of microstructural observations of the distribution of metal sulphide melts in silicate samples that were hydrostatically annealed at high pressures and temperatures^{1–3}. Such observations of melt–solid microstructures are commonly interpreted in terms of the dihedral angle θ , the angle between two neighbouring grains of the solid matrix in contact with melt. This angle is determined by the value of the solid–melt interfacial energy, γ_{sm} , relative to the value of the solid–solid interfacial energy, γ_{ss} , according to the relation:

$$\gamma_{sm}/\gamma_{ss} = 2\cos(\theta/2)$$

For $\theta > 60^\circ$ (that is, large relative values of the solid–melt interfacial energy), the melt does not wet triple junctions, so that percolation is impossible unless the melt fraction exceeds a critical value that increases with increasing dihedral angle⁸. For $0 < \theta < 60^\circ$, the melt phase forms an interconnected network along triple junctions, even at small melt fractions, such that complete drainage of melt by percolation through the solid matrix is possible.

The argument against percolation as a core-forming mechanism is based on the observation that the dihedral angle for metal sulphide melt in the silicate matrix is greater than 60° (refs 1–3). In these high-pressure studies, powders similar in composition to the mantle (that is, olivine with the addition of iron and nickel and iron–nickel sulphides) and powders of a chondritic meteorite believed to be similar in composition to the primitive mantle of Earth were subjected to pressures ranging from 2 to 20 GPa and temperatures of 1,210 to 1,710 °C; these temperatures are higher than the melting point of the metal alloys and metal sulphides, but below the melting point of olivine. Dihedral angles ranged from 70° to 125° ; this shows that the metallic melt is not interconnected, as long as the melt fraction does not exceed a critical value.

For a metallic melt content greater than this critical value in the early Earth, the melt would have been interconnected; melt would have drained to the core until the critical melt fraction was reached in the mantle. At this point, drainage would have ceased, stranding the remaining metallic melt. In theory, the critical melt content is ~ 6 vol.% (for a dihedral angle of 100°)^{4,8}, whereas electrical conductivity experiments place the critical melt content for the olivine–iron system at 10–25% (ref. 9). These values are much larger than the amount of metal stranded in the present-day Earth's mantle¹⁰. In view of the large values measured for the dihedral angle, several authors have concluded that percolation is not a feasible mechanism for the extraction of core-forming melts from the mantle^{1–3}. Consequently, they argue that a significant degree of silicate melting is required to efficiently segregate metal-rich melt from a silicate matrix.

The limitation of this conclusion is, as pointed out by Stevenson⁴, that it is based on experiments carried out under hydrostatic conditions. In contrast, an evolving mantle would not be static,

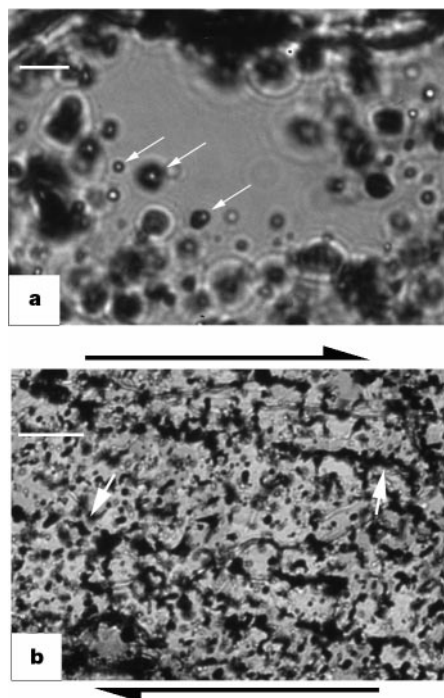


Figure 1 Transmitted light micrographs of olivine + Au samples. Olivine and Au appear light and dark, respectively. **a**, Undeformed sample. The arrows point to spherical Au melt pockets located on an olivine grain boundary. The melt pockets are very small and isolated from each other. Scale bar, 5 μm . **b**, Sample deformed to 250% shear strain. The shear plane is horizontal, and the orientation of the maximum principal stress (σ_1) is 45° to the shear plane such that the direction of shear is top to the right, bottom to the left, as indicated by the large black arrows at the top and bottom of the panel. Trails of Au melt pockets form a pronounced melt-preferred orientation running from top-left to bottom-right. The left-hand white arrow points to a row of aligned but isolated melt pockets, while the right-hand white arrow points to an elongated pocket. Both features are ubiquitous throughout the sample. Scale bar, 10 μm .

but would rather be continually deforming. Hence, the microstructures would reflect dynamic conditions. The shearing motion in a convecting mantle may redistribute the melt, possibly into an interconnected network.

To explore this possibility, and to determine the effect of deformation on the distribution of metal and metal sulphide melts in a crystalline silicate matrix, we performed shear deformation experiments in a high-pressure gas-medium apparatus. Much higher strains can be attained in shear experiments than are possible in deformation experiments with the commonly used triaxial compression technique^{11,12}. We created two types of samples for our experiments. The first was polycrystalline olivine with an elemental gold (Au) melt; this serves as an end-member case for melt connectivity because the dihedral angle is extremely high ($>135^\circ$). The second was polycrystalline olivine with an Fe–Ni–S melt, which approximates the composition of the upper portion of the proto-mantle of Earth.

Both types of samples were made with fine-grained ($<10\text{ }\mu\text{m}$) San Carlos olivine powders. The powders for the olivine-with-Au samples (olivine + 4 vol.% Au powder, $<2\text{ }\mu\text{m}$) were hot-pressed in a gas-medium pressure apparatus at 1,250 °C and 300 MPa for 3 h and quenched at 1°C s^{-1} to below the melting point of Au (1,064 °C). Our olivine-with-metal-sulphide samples (olivine + 6 vol.% FeS and 1 vol.% Ni powder, $\sim 1\text{ }\mu\text{m}$) were hot-pressed in a piston cylinder device at 1 GPa. In a first step, pressure was raised to its final value and then temperature was increased at $20^\circ\text{C min}^{-1}$ to 1,300 °C. At these conditions, the metal sulphide mixture is molten,

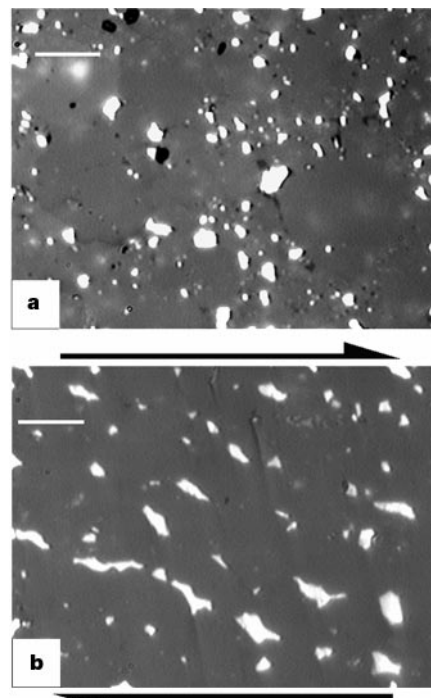


Figure 2 Reflected light photomicrographs of olivine + Fe–Ni–S samples. Olivine and Fe–Ni–S appear light and dark, respectively. Scale bars, 10 μm . **a**, Undeformed sample. The pockets of the metal sulphides are well rounded, and form large dihedral angles with olivine. These large angles indicate that melt pockets are not interconnected so that percolation is not possible. **b**, Sample deformed to 250% shear strain. The shear geometry is identical to that in Fig. 1b. Note that the average melt pocket size increased after shearing, and that melt pockets are aligned at an angle of $\sim 20^\circ$ with the shear plane. The lower left-hand corner is part of the three-dimensional stack shown in Fig. 3.

while olivine forms a solid matrix. After 3 h at temperature and pressure, the temperature was quickly reduced to below the metal sulphide melting temperature, and then slowly decreased ($10^\circ\text{C min}^{-1}$) to room temperature before the pressure was decreased. For both types of samples, the resulting material had less than 1% porosity.

From the resulting aggregates, thin elliptical disks were cut for shear deformation experiments with major and minor axes of 7.5 and 6.1 mm, and thicknesses ranging from 0.6 to 0.9 mm. Shearing was accomplished by placing samples between two thoriated tungsten pistons cut at 45° to the long axis of the pistons. Load was applied parallel to the long axis of the pistons at constant displacement rates of $(2\text{--}3) \times 10^{-5}\text{ mm s}^{-1}$, resulting in shear strain rates of between 10^{-4} s^{-1} and 10^{-3} s^{-1} and shear stresses ranging from 60 to 100 MPa. The experiments were performed at 1,250 °C; confining pressures were 400–450 MPa for the metal sulphide melt samples, and 300–360 MPa for the Au melt samples. To preserve deformation-induced melt microstructures, samples were quenched (at $\sim 1^\circ\text{C s}^{-1}$) under load to below the melting temperature of the metallic phase before the load was reduced.

In undeformed samples, melt pockets are clearly not interconnected (see Figs 1a and 2a). The gold melt has a very high dihedral angle with olivine, and is located in nearly spherical pockets on grain boundaries and in triple junctions. The distribution and topology of the metal sulphide melt is similar to that reported in previous studies^{1–3}.

After shearing our samples to 250% shear strain, we observe very

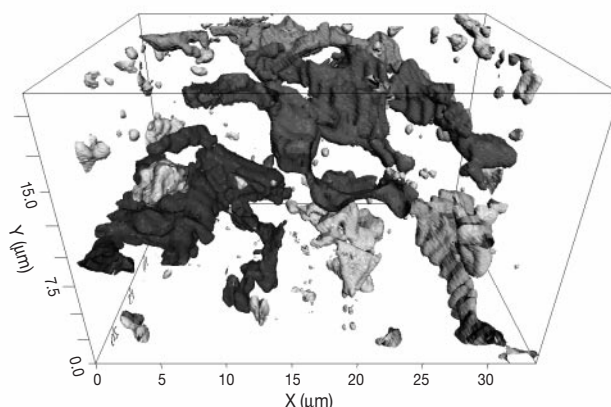


Figure 3 Three-dimensional reconstruction of Fe–Ni–S melt microstructure after deformation. This three-dimensional volume is composed of 40 images such as Fig. 2b, each separated by 0.6 μm in the z -direction. The stack of two-dimensional images was acquired by serial sectioning of the sample, consecutively polishing away 0.6 μm of the sample and taking an image of the same area after each polishing step. The series of

images was then reassembled into a three-dimensional image using image visualization software (IRIS Explorer Version 3.5, NAG Ltd., Oxford, UK, 1996). Pocket size is coded by greyscale, such that darker pockets are larger. Note that most of the melt is concentrated in a small number of large pockets.

different microstructures characterized by a distinct alignment of melt pockets. In the samples containing Au, this melt-preferred orientation forms an angle of 15° from the shear plane, and an angle of $\sim 30^\circ$ with the maximum principal stress, and consists of (1) aligned, elongated (10–20 μm , several grains) melt pockets and (2) rows of isolated pockets oriented 15° from the shear plane (Fig. 1b). Trails of melt containing both of these types of features can be traced for hundreds of micrometres, suggesting that melt pockets have communicated with each other on the scale of the sample during deformation. These trails are part of well-connected sheets of melt extending into the third dimension of the sample, as can be seen in transmitted light with an optical microscope by focusing through the sample.

For the metal sulphide samples, the melt pockets in the deformed sample are more than a factor of two larger than those in the undeformed samples (Fig. 2b). Many of the melt pockets are elongated and oriented similarly to those in the gold melt samples, but at an angle of 20° to the shear plane. In several places, pockets link up to form interconnected features that extend over several tens of micrometres, a distance large compared to the size of the olivine grains. The three-dimensional reconstruction of a series of two-dimensional reflected-light images demonstrates that the aligned melt pockets are well interconnected in the third dimension (Fig. 3).

Similar microstructures have been observed in experimentally sheared aggregates of olivine plus basaltic melt¹³ for which $\theta \approx 30^\circ$ such that the melt was initially interconnected. The alignment of the melt was found to be independent of the crystallographic preferred orientation of olivine, the latter being similar to that observed in melt-free olivine aggregates deformed in shear¹¹. Because deformation produces similar microstructures in olivine aggregates containing different types of melt, we conclude that the formation of a melt-preferred orientation at low angles to the maximum principal stress is independent of both chemical composition and interfacial energy. Instead, the alignment of melt is probably due to the local pressure gradient caused by the applied stress and the deformation geometry. Melt flows from high-pressure to low-pressure regions, assisted by the relative motion of grains past each other. Based on elasticity theory, the alignment of melt observed in our experiments is also expected to exist at the larger grain sizes and lower stresses and strain rates in the Earth's mantle^{13,14}.

The mechanism of the segregation of metal melt from the mantle has been debated ever since Elsasser⁶ proposed a process involving percolation of liquid iron through a solid silicate matrix. Many

recent geophysical (as well as geochemical) models of core formation consider the percolation of metal melt through a solid silicate matrix to be impossible (see, for example, refs 4 and 5). These models therefore require a significant degree of melting of the mantle to make it possible for metal to segregate out of the mantle in order to form the core.

Even though there is no unequivocal geochemical evidence for complete mantle melting on Earth^{16–18}, a number of recent experiments are in good agreement with the existence of a magma ocean^{19,20}. Lunar science observations, and recent numerical models of the thermal evolution of the Earth, also suggest that our planet was molten during most of its accretionary history, and consequently that core formation was accomplished by metal segregation from a magma ocean^{21–23}.

There are several potential mechanisms that could have caused large-scale or complete melting of the Earth. One is accretion of particles to form the planet at high enough temperatures to cause melting (possibly accompanied by radioactive heating); another energy source for melting of a planet is giant impacts²⁴; alternatively, the release of energy due to gravitational sinking of core-forming metals could lead to significant melting of the mantle (see, for example, ref. 25). The last mechanism requires percolation of a metallic melt through a solid silicate matrix: segregation of metallic melts would have been the cause of a magma ocean rather than a consequence.

Even if the Earth's mantle was completely molten before core formation—due either to hot accretion or to the energy released by giant impacts—deformation-assisted percolation of metallic melts may still have been an important core-forming mechanism on other terrestrial planets. Recent experimental and theoretical studies on the partitioning of moderately siderophile elements among silicates and metal sulphide melt, and the comparison of these results with estimates of abundances in the mantle of Mars, suggest²⁶ that the martian core formed in a mostly solid planet. If there was an energy source to drive deformation in the martian mantle, deformation-assisted segregation of metallic melts could explain the existence of a core in Mars without the requirement of extensive silicate melting. A significant amount of deformation of the martian mantle was inferred from the recently reported magnetic lineations in the ancient crust of Mars, which, if confirmed, would suggest that plate-tectonic processes were active in the early history of the planet²⁷. Mantle convection, which is associated with plate tectonics on Earth, also causes large-scale deformation in the planetary

mantle, and would also provide the large shear strains necessary for interconnection and thus percolation of metallic melts in a solid silicate matrix.

Our experimental results show that, in a dynamic environment, melt can be efficiently interconnected even if dihedral angles are large enough to cause the melt to be isolated under hydrostatic conditions. Hence, the requirement of a magma ocean as a critical step in the formation of planetary cores cannot be solely based on the argument that percolation of metal sulphide melt is impossible in a solid silicate mantle. □

Received 30 August; accepted 8 December 1999.

1. Minarik, W. G., Ryerson, F. J. & Watson, E. B. Textural entrapment of core-forming melts. *Science* **272**, 530–533 (1996).
2. Ballhaus, C. & Ellis, D. J. Mobility of core melts during Earth's accretion. *Earth Planet. Sci. Lett.* **143**, 137–145 (1996).
3. Shannon, M. C. & Agee, C. B. High-pressure constraints on percolative core-formation. *Geophys. Res. Lett.* **23**, 2717–2720 (1996).
4. Stevenson, D. J. in *Origin of the Earth* (eds Newsom, H. E. & Jones, J. H.) 231–249 (Oxford Univ. Press, 1990).
5. Karato, S. & Murthy, V. R. Core formation and chemical equilibrium in the Earth – I. Physical considerations. *Phys. Earth Planet. Inter.* **100**, 61–79 (1997).
6. Elssasser, W. M. in *Earth Science and Meteorites* (eds Geiss, J. & Goldberg, E.) 1–30 (North Holland, Amsterdam, 1963).
7. Brett, R. Chemical equilibration of the Earth's core and upper mantle. *Geochim. Cosmochim. Acta* **48**, 1183–1188 (1984).
8. Von Bagen, N. & Waff, H. S. Permeabilities, interfacial areas and cutrvatures of partially molten systems: Results of numerical computations of equilibrium microstructures. *J. Geophys. Res.* **91**, 9261–9276 (1986).
9. Jurewicz, S. R. & Jones, J. H. Preliminary results of olivine/metal wetting experiments and the direct measurement of metal phase interconnectivity. *Proc. Lunar Planet. Sci. Conf.* **26**, 709–710 (1995).
10. O'Neill, H. St. C. & Palme, H. in *The Earth's Mantle* (ed. Jackson, I.) 3–123 (Cambridge Univ. Press, 1998).
11. Zhang, S. & Karato, S. Lattice-preferred orientation of olivine aggregates deformed in simple shear. *Nature* **375**, 774–777 (1995).
12. Karato, S., Zhang, S., Zimmerman, M. E., Daines, M. J. & Kohlstedt, D. L. Experimental studies of shear deformation of mantle materials; towards structural geology of the mantle. *Pure Appl. Geophys.* **151**, 589–603 (1998).
13. Zimmerman, M. E., Zhang, S., Kohlstedt, D. L. & Karato, S. Melt distribution in mantle rocks deformed in shear. *Geophys. Res. Lett.* **26**, 1505–1508 (1999).
14. Daines, M. J. & Kohlstedt, D. L. Influence of deformation on melt topology in peridotites. *J. Geophys. Res.* **107**, 10257–10271 (1997).
15. Arculus, R. J., Holmes, R. D., Powell, R. & Righter, K. in *Origin of the Earth* (ed. Newsom, H. E. & Jones, J. H.) 251–272 (Oxford Univ. Press, 1990).
16. Walter, M. At the magma ocean floor. *Nature* **381**, 646 (1996).
17. Stevenson, D. J. in *Workshop on the Early Earth: The Interval from Accretion to the Older Archean* (eds Burke, K. & Ashwal, L. D.) 76–78 (LPI Tech. Rep. 85-01, Lunar and Planetary Institute, Houston, 1985).
18. Ringwood, A. E. in *Origin of the Earth* (eds Newsom, H. E. & Jones, J. H.) 101–134 (Oxford Univ. Press, 1990).
19. Li, J. & Agee, C. B. Geochemistry of mantle-core formation at high pressure. *Nature* **381**, 686–689 (1996).
20. Righter, K. & Drake, M. J. Effect of water on metal-silicate partitioning of siderophile elements: a high pressure and temperature terrestrial magma ocean and core formation. *Earth Planet. Sci. Lett.* **171**, 383–399 (1999).
21. Murthy, V. R. Early differentiation of the Earth and the problem of mantle siderophile elements: A new approach. *Science* **253**, 303–306 (1991).
22. Benz, W. & Cameron, A. G. W. in *Origin of the Earth* (eds Newsom, H. E. & Jones, J. H.) 61–67 (Oxford Univ. Press, 1990).
23. Agee, C. B. & Longhi, J. (eds) *Workshop on the Physics and Chemistry of Magma Oceans from 1 bar to 4 Mbar* (LPI Tech. Rep. 92–03, Lunar and Planetary Institute, Houston, 1992).
24. Wetherill, G. W. Formation of the Earth. *Annu. Rev. Earth Planet. Sci.* **18**, 205–256 (1990).
25. Hanks, T. C. & Anderson, D. L. The early thermal history of the Earth. *Phys. Earth Planet. Inter.* **2**, 19–29 (1969).
26. Gaetani, G. A. & Grove, T. L. Partitioning of moderately siderophile elements among olivine, silicate melt, and sulphide melt: Constraints on core formation in the Earth and Mars. *Geochim. Cosmochim. Acta* **61**, 1829–1846 (1997).
27. Connerney, J. E. P. et al. Magnetic lineations in the ancient crust of Mars. *Science* **284**, 794–798 (1999).

Acknowledgements

We thank D. Xirouchakis for help with fabrication of our samples, and M. Hirschmann, R. Murthy and S. Karato for discussion and comments on an earlier version of the manuscript. W. Minarik and M. Walter provided constructive comments on the manuscript. P. Morin produced the three-dimensional visualization of our deformation microstructures (Fig. 3). This work was supported by Deutsche Forschungsgemeinschaft (D.B.), NASA and the NSF.

Correspondence and requests for materials should be addressed to D.B. (e-mail: bruhn@olivine.geology.umn.edu).

Rapid evolution of reproductive barriers driven by sexual conflict

Sergey Gavrillets

Departments of Ecology and Evolutionary Biology and Mathematics,
University of Tennessee, Knoxville, Tennessee 37996-1610, USA

A growing amount of experimental data indicates extremely rapid evolution of traits and proteins related to fertilization in many diverging taxa^{1–3}. These data come from studies of sperm or pollen competition between closely related species^{3–6}, and from molecular studies of fertilization proteins^{2,7–10}. The positive selection for evolutionary novelty that appears to be acting on fertilization systems seems paradoxical because successful reproduction requires the close matching of female and male traits. It has been suggested^{11–13} that perpetual coevolution between the sexes can result from sexual conflict in mating. Sexual conflict occurs when characteristics that enhance the reproductive success of one sex reduce the fitness of the other sex¹⁴. Numerous examples of sexual conflict resulting from sensory exploitation, polyspermy and the cost of mating have been discussed in detail^{1–3,14,15}. The potential for coevolution due to such conflict has been evaluated experimentally^{15,16}. Here I develop a simple mathematical model describing coevolutionary dynamics of male and female traits involved in reproduction. The model shows that continual change in such traits at a constant speed is expected whenever females (or eggs) experience fitness loss from having too many compatible males (or sperms). The plausibility of runaway coevolution increases with increasing population size. Rapid evolution of reproductive barriers driven by sexual conflict may explain increased speciation rates after colonization of new habitats ('adaptive radiation') and high species richness in resource-rich environments.

An important population characteristic that can be used to incorporate different forms of sexual conflict into the modelling framework is the proportion of the individuals (or gametes) of the other sex that can mate with an individual (or gamete) of a given type. An implicit assumption of classical population genetic models (which did not consider sexual conflicts) is that individual fitness maximizes when this proportion is one. Here I study the coevolutionary dynamics of male and female traits when sexual conflict over mating rate results in an optimal proportion of compatible males for females being smaller than one.

I will say that two individuals (or gametes) are 'compatible' if mating and fertilization are not prevented by isolating mechanisms. Let the degree of 'compatibility' of males and females (or their gametes) be controlled by two independent sex-limited quantitative traits: a male (or sperm) trait y and a female (or egg) trait x . This independence assumption is justified at least for gamete recognition systems². Let $f(x)$ and $g(y)$ be the distributions of x and y in the population with means $\bar{x}$ and $\bar{y}$ and variances V_x and V_y . I assume that compatibility of sexes requires a matching of the corresponding traits. The simplest choice for a corresponding function defining the probability $\Psi(x, y)$ that trait x is compatible with trait y is a quadratic function:

$$\Psi(x, y) = 1 - \alpha(x - y)^2$$

where α is a positive parameter controlling the tolerance of matching¹⁷. The proportion of the males compatible with female trait x is

$$P(x) = \int \Psi(x, y)g(y)dy = 1 - \alpha[V_y + (x - \bar{y})^2]$$

The overall fitness $W_f(x)$ of females (or eggs) having trait x depends

on $P(x)$. This fitness should be an increasing function at small values of $P(x)$ (the more compatible males the higher the probability of fertilization), but should decrease at larger values of $P(x)$ (as a result of polyspermy, predation, sensory exploitation, different costs of mating, seminal fluid toxicity and other forms of sexual conflict; see ref. 18 for a meta-analysis of relevant insect data). The simplest form of a function having these properties is a quadratic function with a maximum at P_{opt} :

$$W_f(x) = 1 - S[P(x) - P_{\text{opt}}]^2 \quad (1)$$

where S is a positive parameter. Instead of parameters S and P_{opt} it is more illuminating to use parameters s and θ , which are defined by the equalities $P_{\text{opt}} = 1 - \theta$ and $S = s/\theta^2$. That is, parameter θ is the amount by which the optimal proportion of compatible males is smaller than 1, and parameter s is the fitness reduction of a female trait that is compatible with all males (for which $P(x) = 1$). Note that choosing a negative or zero value for θ would describe a situation in which female fitness always increases with $P(x)$ (no sexual conflict). The probability that a male (or sperm) y has offspring from mating a female (or egg) x is a product of the probability $\Psi(x, y)$ that traits x and y are compatible and the probability $F(x)$ that fertilization does take place and result in viable offspring. If there are too many males compatible with female trait x , the chance to be the one who fertilizes a given female decreases. Moreover, even if fertilization does take place, interference from other males (for example, through polyspermy, seminal fluid toxicity, physical harm to females, costs of mating) can prevent successful development of offspring^{1,11,12}. Therefore, probability F should be a decreasing function of the overall proportion $P(x)$ of males that are compatible with female trait x . The simplest form of a function having this property is a linear function: $F = 1 - BP$, where $0 < B < 1$. To find the male fitness $W_m(y)$ one sums over all females:

$$W_m(y) = \int \Psi(x, y) F(x) f(x) dx \quad (2)$$

Note that both male and female trait fitnesses (defined by equations (1) and (2), respectively) are frequency dependent.

I will assume that $\alpha \ll \theta$, $s \ll 1$. These assumptions imply that selection is weak and linkage disequilibrium can be neglected. I will allow the term $\alpha(\bar{x} - \bar{y})^2$ to have the same order of magnitude as θ and s . Under the above assumptions, the per generation changes in the means are approximately

$$\Delta \bar{x} = 2\alpha V_x \frac{s}{\theta} \left[1 - \frac{\alpha(\bar{x} - \bar{y})^2}{\theta} \right] (\bar{x} - \bar{y}) \quad (3a)$$

$$\Delta \bar{y} = \alpha V_y (\bar{x} - \bar{y}) \quad (3b)$$

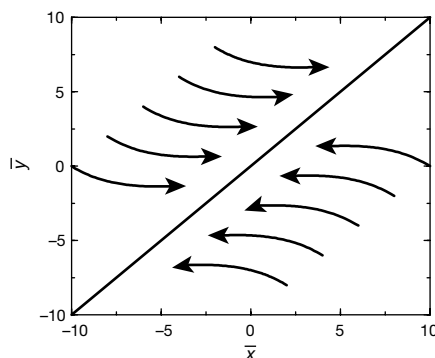


Figure 1 Evolution towards a stable line of equilibria $\bar{y} = \bar{x}$. Parameter values used are $\alpha = 0.01$, $V_x = V_y = 1$, $s = 0.01$, $\theta = 0.15$. The line of equilibria and a set of ten trajectories (corresponding to 100 generations each) starting from different initial states are shown. The direction of change is indicated by arrows.

Equation (3b) shows that if no female trait evolution is allowed (that is, if $\bar{x}$ is fixed, say, because there is no genetic variation, $V_x = 0$, or by some experimental manipulation¹⁵) the mean male trait evolves to a value matching $\bar{x}$. One can say that for the male trait the optimal strategy is to match the mean female trait. When female *D. melanogaster* are experimentally prevented from coevolving with males, males rapidly adapt to the static female phenotype, and this male adaptation leads to a reduction in female survival¹⁵. Alternatively, if no male trait evolution is allowed (that is, if $\bar{y}$ is fixed) equation (3a) shows that the mean female trait will evolve to a value that differs from $\bar{y}$ by $\pm(\theta/\alpha)^{1/2}$. One can say that this value represents a displacement of the male and female traits that is optimal for females. The difference between the displacements that are optimal for males and females is a result of the sexual conflict over mating rate.

To analyse the dynamics when both male and female traits are allowed to evolve, let us first make a standard assumption that genetic variances V_x and V_y are constant (for example, see refs 17, 19, 20). Then there are two possible dynamical regimes. In the first regime, the system evolves to an equilibrium state at which the mean trait values match: $\bar{y} - \bar{x} \rightarrow 0$ (Fig. 1). One can say that in this case males win the sexual conflict. The dynamical system (3) has a stable line of equilibria similar to those arising in models of sexual selection¹⁷. Changes in the trait values along the line $\bar{y} = \bar{x}$ are neutral, and different allopatric populations can diverge along this line by random genetic drift (see ref. 17). The rate of such a divergence will increase with decreasing the population size. This regime takes place if

$$1 < \frac{1}{2} \frac{V_y \theta}{V_x s} \quad (4)$$

The ratio of the variances in the righthand side of equation (4) characterizes the relative ability of male and female traits to evolve. The second ratio characterizes the strength of the sexual conflict: the smaller the fitness loss s is at a given value θ , the less intense the sexual conflict. Thus, the first regime is promoted if the genetic variation in males is higher than in females, or if the fitness loss of females from having too many compatible males is small.

If inequality (4) is reversed, the difference in the means approaches asymptotically a constant value displaced from zero: $\bar{y} - \bar{x} \rightarrow \pm \delta$, where

$$\delta = \sqrt{\frac{\theta}{\alpha} \left(1 - \frac{1}{2} \frac{V_y \theta}{V_x s} \right)} \quad (5)$$

The value of the displacement δ is intermediate between those that are optimal for males (zero displacement) and females (displacement

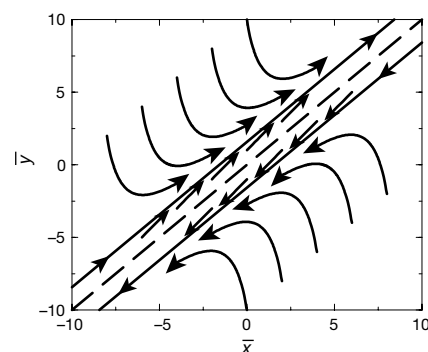


Figure 2 Runaway dynamics along lines $\bar{y} = \bar{x} \pm \delta$. Parameter values used are $\alpha = 0.01$, $V_x = V_y = 1$, $s = 0.1$, $\theta = 0.15$. A line of unstable equilibria $\bar{y} = \bar{x}$ (dashed line), the lines of runaway coevolution $\bar{y} = \bar{x} + \delta$ and $\bar{y} = \bar{x} - \delta$ and a set of 20 trajectories (corresponding to 100 generations each) starting from different initial states are shown. The direction of change is indicated by arrows.

at a value of $(\theta/\alpha)^{1/2}$. Thus, neither sex wins the sexual conflict but rather there is a coevolutionary compromise. This compromise is dynamic—neither male nor female traits settle to an equilibrium but both keep simultaneously increasing (or decreasing) along a line at which the displacement is constant. In sharp contrast to the models in which runaway coevolution occurs at a geometrically increasing rate^{17,20} or is cyclic^{19,20}, here the rate R of morphological changes (that is the change in the means per generation $R = \Delta\bar{x} = \Delta\bar{y}$) is constant:

$$R = \alpha V_y \delta \quad (6)$$

The resulting dynamics are illustrated in Fig. 2.

In both regimes, the characteristic time for approaching a line $\bar{y} - \bar{x} = \text{const}$ is about $1/[\alpha(2V_x(s/\theta) - V_y)]$ generations. For example, let $\alpha = 0.01$, $V_x = V_y = 1$, $s = 0.1$, $\theta = 0.15$. Then depending on the initial conditions, the system approaches one of the two lines: $\bar{y} = \bar{x} + \delta$, or $\bar{y} = \bar{x} - \delta$, where $\delta \approx 1.58$. The characteristic approach time is about 300 generations. The rate of runaway coevolution along the line is $R = 0.0158$. That is, it takes about 63 generations for a shift of the order of one standard deviation. Thus, a few hundred generations will be sufficient for a significant change in the mean trait values.

The ratio of the variances V_y and V_x in inequality (4) is a crucial parameter affecting the qualitative behaviour of the system. On a short time interval this ratio can be treated as a constant (numerically) close to its initial value. On larger time scales, however, genetic variances can change because they are affected by both selection and mutation. Although modelling the dynamics of genetic variances is notoriously difficult^{21,22}, some progress can be made. Let mutation increase genetic variances V_x and V_y by μ_x and μ_y per generation, respectively. Using the method that is consistent with the derivation of equation (3) results in a prediction that at a mutation–selection balance equilibrium

$$V_y^* = \sqrt{\mu_y/\alpha}, \quad (7a)$$

$$V_x^* = \frac{3}{8} V_y^* \frac{\theta}{s} \left(1 + \sqrt{1 + \frac{16\mu_x s}{9\mu_y \theta}} \right) \quad (7b)$$

At this equilibrium, inequality (4) is never true. Thus, mutation and selection are expected to modify genetic variances in such a way that the system always enters the regime in which the means evolve in a runaway fashion. The characteristic time for reaching a mutation–selection balance state is on the order of $1/(\mu\alpha)^{1/2}$ generations. With $\mu_x = \mu_y = 0.01$ and with the same values of α , s and θ as above, the levels of genetic variation maintained are relatively high: $V_x = 1.39$, $V_y = 1$. Studies of gamete recognition proteins indicate that such genes can be highly polymorphic within species². The characteristic time for approaching a mutation–selection balance state is about 100 generations. The above parameter values result in $\delta = 2.15$ and $R = 0.0215$.

A main conclusion of the model studied here is that sexual conflict naturally leads to perpetual changes in traits responsible for reproductive isolation. The sexes get locked in coevolutionary chase in which females continuously evolve to decrease the mating rate while males continuously evolve to increase it. This theoretical conclusion supports previous verbal arguments^{11–13}. What was not expected is the generality of conditions under which coevolutionary chase is expected to occur: continual change in traits responsible for reproductive barriers at a constant speed is expected whenever females (or eggs) experience fitness loss from having too many compatible males (or sperms). Allopatric populations can evolve in the opposite directions or in the same direction but at different rates. In either case, the expected outcome will be rapid genetic divergence and eventual reproductive incompatibility of individuals (or gametes) from different populations. This conclusion is sup-

ported by data²³ that indicate a striking increase in the rate of speciation in groups in which post-mating sexual conflict is intense as compared with related groups in which it is absent.

In earlier models of allopatric speciation driven by genetic drift along a line of equilibria arising under sexual selection¹⁷, or by the accumulation of chromosome rearrangements²⁴, or by mutation and drift on holey adaptive landscapes^{25,26}, the rate of evolution of reproductive barriers is the fastest in small populations. In contrast, because sexual conflict induces direct selection on traits involved in reproduction, large populations will be more responsive to such selection. There is an additional effect of population size on the evolutionary dynamics driven by sexual conflict. Increasing the overall population sizes should intensify sexual conflict (see ref. 27 for relevant examples on water striders). In terms of this model, this translates into the optimum proportion of compatible males becoming smaller than one (θ becomes positive) and an increasing fitness loss of females from having too many compatible males (s becomes bigger). Both these effects will make runaway coevolution of sexes more plausible. Thus, increasing the sizes of geographically isolated populations should make evolution of reproductive barriers between them more plausible. In large populations, significant changes in traits that are involved in reproduction may occur on the timescale of a few hundred to a few thousand generations.

Two explanations of increased speciation rates ('adaptive radiation') after colonization of new resource-rich environments free of competitors and/or predators have been intensively discussed²⁸. One is strong divergent natural selection that causes rapid phenotypic differentiation and, as a by-product, reproductive isolation. Another is increased persistence of local populations that increases the chance that they avoid extinction long enough to evolve reproductive isolation. Our model provides a third possibility: increased speciation rates may be caused by sexual conflict that intensifies after increasing the size of local populations. Founding several isolated populations that quickly increase in size should result in rapid divergence of these populations in traits involved in reproduction. The same argument can be used to explain high species richness that is observed in geographic areas with high productivity²⁹. With sexual conflict, high local population densities that can be maintained in such areas should promote the evolution of reproductive barriers and high species numbers. One way to test this is by comparing the degree of reproductive isolation between natural or laboratory populations that have been maintained for some time at different densities. The hypothesis that rapid speciation is driven by sexual conflict predicts that reproductive barriers will be stronger between populations with high densities than between populations with low densities. Another prediction is that the number of species in a clade should positively correlate with species abundance. It remains to be seen whether these hypotheses are supported by data. □

Methods

Let $p(z)$ be the distribution of a quantitative trait z in the population with the mean $\bar{z}$ and central moments $M_i (= \int (z - \bar{z})^i p(z) dz)$. The mean $\bar{z}'$ and the variance M_2' of the trait after selection specified by fitness function $w(z)$ are

$$\bar{z}' = \frac{\int zw(z)p(z)dz}{\bar{w}} \quad (8a)$$

$$M_2' = \frac{\int z^2 w(z)p(z)dz}{\bar{w}} - (\bar{z}')^2, \quad (8b)$$

where $\bar{w} = \int w(z)p(z)dz$ is the mean fitness. Let the fitness function be represented as a polynomial in $z - \bar{z}$

$$w(z) = a_0 + \sum_{i=1}^k a_i (z - \bar{z})^i, \quad (9)$$

where coefficients a_i are allowed to depend on the moments of $p(z)$ but not on z . Then the within generation changes in the mean and variance are

$$\Delta \bar{z} = \frac{\sum_{i=1}^k a_i M_{i+1}}{a_0 + \sum_{i=2}^k a_i M_i} \quad (10a)$$

$$\Delta M_2 = \frac{\sum_{i=1}^k a_i (M_{i+2} - M_i M_2)}{a_0 + \sum_{i=2}^k a_i M_i} - (\Delta \bar{z})^2 \quad (10b)$$

(compare refs 19–21). These equations are exact. If trait z is additive and linkage disequilibrium can be neglected, segregation and recombination will not change $\bar{z}$ and M_2 . Thus, equations (10) will give the changes in $\bar{z}$ and M_2 between two subsequent generations. For traits with sex-limited expression (such as those considered here) the changes in $\bar{z}$ and M_2 between two subsequent generations will be equal to one-half of the values predicted by equations (10). The fitness functions for male and female traits defined in the main body of the paper belong to a class of polynomial fitness functions (9) with $k = 4$ and $k = 2$, respectively. Resulting equations for the means are given in the main body of the paper whereas the per generation changes in genetic variances are

$$\Delta V_x = 2\alpha V_x^2 \left[1 - 3 \frac{\alpha(\bar{y} - \bar{x})^2}{\theta} \right] + \mu_x \quad (11a)$$

$$\Delta V_y = -\alpha V_y^2 + \mu_y \quad (11b)$$

where μ_x and μ_y are genetic variances introduced by mutation. To derive these equations (see Supplementary Information for details) one identifies coefficients a_i of the polynomial expansion (9), then plugs them into equations (10), and simplifies the resulting expressions by assuming that $\alpha \ll \theta$, $s \ll 1$, and allowing the term $\alpha(\bar{x} - \bar{y})^2$ to have the same order of magnitude as θ and s . In addition, to derive the equations for changes in the means, I neglected the third moments; and to derive the equations for the variances, I neglected the third moments and assumed that $M_4 - 3M_2^2 = 0$ (zero kurtosis). Numerical results for a case of stabilizing selection with moving optimum³⁰ suggest that this is a satisfactory approximation. The dynamic equations (3) and (11) can be analysed by standard methods.

Received 13 September; accepted 7 December 1999.

1. Rice, W. R. in *Endless Forms. Species and Speciation* (eds Howard, D. J. & Berlocher, S. H.) 261–270 (Oxford Univ. Press, New York, 1998).
2. Palumbi, S. R. in *Endless Forms. Species and Speciation* (eds Howard, D. J. & Berlocher, S. H.) 271–278 (Oxford Univ. Press, New York, 1998).
3. Howard, D. J., Reece, M., Gregory, P. G., Chu, J. & Cain, M. L. in *Endless Forms. Species and Speciation* (eds Howard, D. J. & Berlocher, S. H.) 279–288 (Oxford Univ. Press, New York, 1998).
4. Arnold, M. L., Hamrick, J. L. & Bennett, B. D. Interspecific pollen competition and reproductive isolation in *Iris*. *J. Hered.* **84**, 13–16 (1993).
5. Wade, M. J., Patterson, H., Chang, N. W. & Johnson, N. A. Postcopulatory, prezygotic isolation in flour beetles. *Heredity* **72**, 163–167 (1994).
6. Rieseberg, L. H., Desrochers, A. M. & Youn, S. J. Interspecific pollen competition as a reproductive barrier between sympatric species of *Helianthus* (Asteraceae). *Am. J. Bot.* **82**, 515–519 (1995).
7. Aguade, M., Miyashita, N. & Langley, C. H. Polymorphism and divergence of the mst 355 male accessory gland region. *Genetics* **132**, 755–770 (1992).
8. Palumbi, S. R. Marine speciation on a small planet. *Trends Ecol. Evol.* **7**, 114–117 (1992).
9. Vacquire, V. D. & Lee, Y.-H. Abalone sperm lysin: unusual mode of evolution of a gamete recognition protein. *Zygote* **1**, 181–196 (1993).
10. Metz, E. C. & Palumbi, S. R. Positive selection and sequence rearrangements generate extensive polymorphism in the gamete recognition protein bindin. *Mol. Biol. Evol.* **13**, 397–406 (1996).
11. Rice, W. R. & Holland, B. The enemies within: intergenomic conflict, interlocus contest evolution (ICE) and the intraspecific red queen. *Behav. Ecol. Sociobiol.* **41**, 1–10 (1997).
12. Holland, B. & Rice, W. R. Chase-away sexual selection: antagonistic seduction versus resistance. *Evolution* **52**, 1–7 (1998).
13. Parker, G. A. & Partridge, L. Sexual conflict and speciation. *Phil. Trans. R. Soc. Lond. B* **353**, 261–274 (1998).
14. Stockley, P. Sexual conflict resulting from adaptations to sperm competition. *Trends Ecol. Evol.* **12**, 154–159 (1997).
15. Rice, W. R. Sexually antagonistic male adaptation triggered by experimental arrest of female evolution. *Nature* **381**, 232–234 (1996).
16. Holland, B. & Rice, W. R. Experimental removal of sexual selection reverses intersexual antagonistic coevolution and removes a reproductive load. *Proc. Natl Acad. Sci. USA* **96**, 5083–5088.
17. Lande, R. Models of speciation by sexual selection on polygenic characters. *Proc. Natl Acad. Sci. USA* **78**, 3721–3725 (1981).
18. Arnqvist, G. & Nilson, T. The evolution of polyandry: multiple mating and female fitness in insects. *Anim. Behav.* (submitted).
19. Iwasa, I. & Pomiankowski, A. Continual change in mate preferences. *Nature* **377**, 420–422 (1995).
20. Gavrillets, S. Coevolutionary chase in exploiter–victim systems with polygenic characters. *J. Theor. Biol.* **186**, 527–534 (1997).
21. Barton, N. H. & Turelli, M. Natural and sexual selection on many loci. *Genetics* **127**, 229–255 (1991).
22. Gavrillets, S. & Hastings, A. Dynamics of polygenic variability under stabilizing selection, recombination and drift. *Genet. Res.* **65**, 63–74 (1995).
23. Arnqvist, G., Edvardsson, M., Friberg, U. & Nilsson, T. Sexual conflict promotes speciation in insects. *Proc. R. Soc. Lond. B* (submitted).
24. Walsh, J. B. Rate of accumulation of reproductive isolation by chromosome rearrangements. *Am. Nat.* **120**, 510–532 (1982).
25. Gavrillets, S. Evolution and speciation on holey adaptive landscapes. *Trends Ecol. Evol.* **12**, 307–312 (1997).
26. Gavrillets, S. A dynamical theory of speciation on holey adaptive landscapes. *Am. Nat.* **154**, 1–22 (1999).

27. Arnqvist, G. in *The Evolution of Mating Systems in Insects and Arachnids* (eds Choe, J. C. & Crespi, B. J.) 146–163 (Cambridge Univ. Press, Cambridge, Massachusetts, 1997).
28. Schluter, D. in *Endless Forms. Species and Speciation* (eds Howard, D. J. & Berlocher, S. H.) 114–129 (Oxford Univ. Press, New York, 1998).
29. Rosenzweig, M. L. *Species Diversity in Space and Time* (Cambridge Univ. Press, Cambridge, Massachusetts, 1995).
30. Bürger, R. & Lynch, M. Evolution and extinction in a changing environment—a quantitative genetic analysis. *Evolution* **49**, 151–163.

Supplementary information is available on Nature's World-Wide Web Site (<http://www.nature.com>) or as paper copy from the London editorial offices of Nature.

Acknowledgements

I am grateful to G. Arnqvist, C. R. B. Boake, M. B. Cruzan and G. McCracken for discussions and/or helpful comments on the manuscript. This work was supported by an NIH grant and by a grant from the Netherlands Organization for Scientific Research (NWO).

Correspondence and requests for materials should be addressed to S.G. (e-mail: sergey@tiem.utk.edu).

Thermal stimulation of taste

Alberto Cruz* & Barry G. Green*†

* The John B. Pierce Laboratory and † Department of Surgery (Otolaryngology), Yale School of Medicine, 290 Congress Avenue, New Haven, Connecticut 06519, USA

The first electrophysiological recordings from animal¹ and human² taste nerves gave clear evidence of thermal sensitivity, and studies have shown that as many as half of the neurons in mammalian taste pathways respond to temperature^{3–6}. Because temperature has never been shown to induce sensations of taste, it has been assumed that thermal stimulation in the gustatory system is somehow nulled⁶. Here we show that heating or cooling small areas of the tongue can in fact cause sensations of taste: warming the anterior edge of the tongue (chorda tympani nerve) from a cold temperature can evoke sweetness, whereas cooling can evoke sourness and/or saltiness. Thermal taste also occurs on the rear of the tongue (glossopharyngeal nerve), but the relationship between temperature and taste is different there than on the front of the tongue. These observations indicate the human gustatory system contains several different types of thermally sensitive neurons that normally contribute to the sensory code for taste.

Research into thermal effects on taste has focused on modulation of sensitivity to the flavours of chemicals (see, for example, refs 7–9) rather than on the possibility that temperature itself might stimulate taste. Our attention was drawn to this possibility during preliminary experiments on the thermal sensitivity of the tongue. We noticed that warming the tongue tip from 20 to 35 °C caused a transient sensation of sweetness, and that cooling it to ≤20 °C induced a sour taste that for one of us turned to saltiness at temperatures below 10 °C. A screening test for thermal taste conducted on 24 naive subjects yielded 21 individuals who reported at least one taste quality and 19 (5 males and 14 females) who reliably reported two or more tastes at one or more sites along the anterior edge of the tongue. Tastes were generally described as weak but expressions of surprise at their clarity and strength were not uncommon. The subjects who reported two or more taste qualities were enrolled in a study of the psychophysical properties of thermal taste.

We began by exploring the relationship between temperature and taste quality. Figure 1a shows that warming the tongue tip from 20 °C induced sweetness but not other significant tastes in 16 subjects (repeated measures analysis of variance (ANOVA) (temperature × taste quality × replicate), main effect of quality; $F(3, 45) = 40.18$, $P < 0.05$). Sweetness intensified between tem-

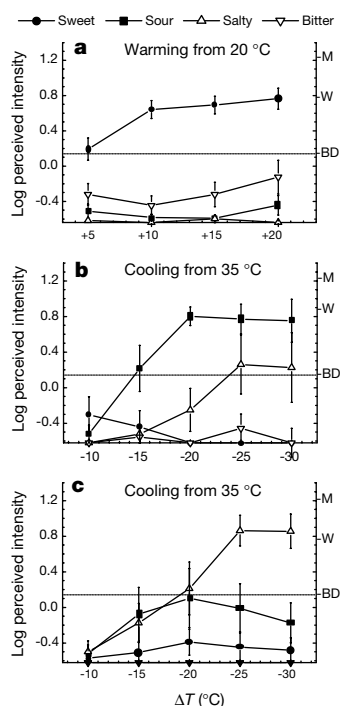


Figure 1 The perceived intensity of taste sensations reported during thermal stimulation of the midline of the tongue tip. The log of the mean perceived intensity is plotted against ΔT , the temperature change. On warming trials (**a**) temperature was increased from 20 °C; on cooling trials (**b, c**) temperature was decreased from 35 °C. Each graph contains the data from those subjects (**a**, $n = 16$; **b**, $n = 6$; **c**, $n = 6$) whose average rating of the dominant taste quality (sweetness, sourness, saltiness) exceeded 'barely detectable' on the labelled magnitude scale¹⁶ (LMS). Letters along the right y-axis on these and subsequent graphs denote intensity descriptors on the LMS: BD, barely detectable (horizontal dotted line); W, weak; M, moderate. Vertical bars indicate standard errors of the means (s.e.m.s).

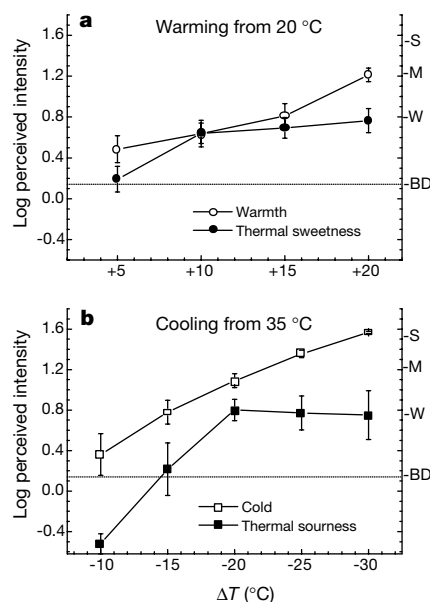


Figure 2 Temperature and thermal taste ratings as a function of temperature change at the tongue tip. **a**, During warming; **b**, During cooling. The thermal taste data are from Fig. 1a and b; the temperature data are from the same subjects.

perature changes (ΔT s) of +5 and +10 °C (temperature $\times$ taste quality interaction, $F(9, 135) = 4.96$, $P < 0.05$; Tukey HSD test, $P < 0.05$), but further warming had little additional effect. Figure 1b shows that cooling the tongue tip by 15 to 20 °C evoked sourness in 6 subjects (main effect of quality, $F(12, 60) = 4.61$, $P < 0.05$) which did not intensify with further cooling. Another 6 subjects (Fig. 1c), 3 of whom had reported sourness at ΔT s of -15 and -20 °C, rated saltiness as the dominant quality at colder temperatures (main effect of quality, $F(3, 15) = 6.31$, $P < 0.05$; interaction between temperature and quality, $F(12, 60) = 6.24$, $P < 0.05$). In contrast to thermal taste, perceived warmth and cold both continued to intensify as ΔT increased (Fig. 2).

Next we studied the sites on the edge of the tongue where the two tastes most frequently reported during the screening test, thermal sweetness (T_{SW}) and sourness (T_{SO}), were maximal. 'Best sites' for both taste qualities identified from the screening data indicated the locations of peak sensitivity to T_{SW} and T_{SO} did not coincide: the best sites for thermal sweetness ('sweet-best sites') were always near the tongue tip, whereas all but one of the 'sour-best sites' were lateral to the tip. Indeed, when testing was not restricted to the tongue tip, the number of subjects who reliably reported sourness rose from 6 to 15. Systematic measurements on sweet-best and sour-best sites showed that warming from 20 to 35 °C failed to stimulate significant T_{SW} on sour-best sites (Fig. 3a, left), and cooling from 35 to 15 °C evoked only barely detectable T_{SO} on sweet-best sites (Fig. 3b, left). Cooling to 5 °C tended to evoke T_{SO} on sour-best sites and thermal saltiness (T_{SA}) on sweet-best sites, but this difference was not significant (Fig. 3c, left). Ratings of warmth and cold revealed no difference in responsiveness to cooling between sweet- and sour-best sites, but warmth was rated significantly more intense on sweet-best sites (t -test for dependent samples $t_{18} = 3.15$, $P < 0.01$).

The same sites were tested with chemical stimuli to study the

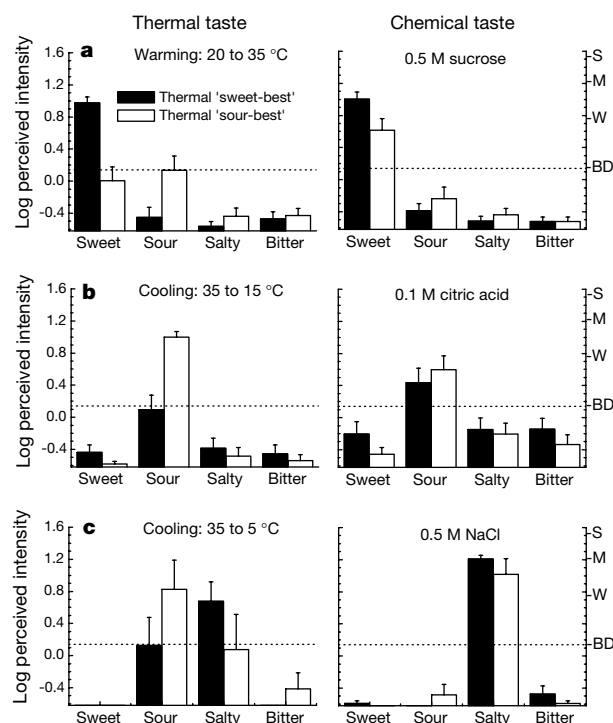


Figure 3 Perceived intensity of thermal taste (left column) and chemical taste (right column) on different sites on the tongue. Data are shown for the 'best' sites for thermal sweetness (filled bars) and thermal sourness (open bars). The data shown are again from those subjects who reported greater than 'barely detectable' sweetness (**a**, $n = 18$), sourness (**b**, $n = 15$) or saltiness (**c**, $n = 5$) during warming (20 to 35 °C), moderate cooling (35 to 15 °C), or extreme cooling (35 to 5 °C), respectively.

relationship between thermal and chemical taste. The graphs on the right side of Fig. 3 show that whereas sweet- and sour-best sites were both sensitive to sucrose, citric acid and NaCl, sweet-best sites (Fig. 3a, right) yielded higher sweetness ratings for sucrose than did sour-best sites ($t_{18} = 2.84$, $P < 0.02$). Tendencies for citric acid to be rated more sour on sour-best sites and NaCl to be rated saltier on sweet-best sites did not reach significance. Bitterness (quinine hydrochloride (QHCL); not shown) was perceived equally on both sites. While these results point to a spatial relationship between thermally and chemically induced sweetness, they also indicate that sensitivity to thermal taste is not distributed as uniformly as sensitivity to chemical taste.

The finding that sweet-best sites yielded higher warmth ratings than sour-best sites prompted a more detailed survey of the association between thermal taste and thermal sensitivity along the anterior edge of the tongue. Stepwise spatial testing confirmed that T_{SW} was strongest near the tip of the tongue (repeated measures ANOVA, main effect of stimulus location, $F(6, 78) = 38.1$, $P < 0.05$) and was closely associated with perception of warmth (Fig. 4a). In contrast, T_{SO} was no stronger on the tongue tip than it was more laterally (Fig. 4b), and a significant stimulus $\times$ location interaction ($F(6, 72) = 2.47$, $P < 0.05$) confirmed that its spatial pattern differed from cold. Although in Fig. 4 T_{SW} appears stronger on average than T_{SO} , five of the 13 subjects who reported sourness rated it above 'moderate' on one or more test sites. The lower T_{SO} ratings were therefore caused in part by greater variation in the location of sour-best sites compared to sweetness-best sites.

The striking spatial coincidence of warmth and T_{SW} implies that C-warm fibre receptors are among the trigeminal nerve fibres that terminate in fungiform papillae¹⁰, and raises the question whether certain types of trigeminal and chorda tympani neurons may become spatially associated during development and innervation of these papillae¹¹.

We also studied thermal taste on circumvallate papillae, which are innervated by the glossopharyngeal nerve. Twelve subjects from the first experiment who could place the temperature stimulator on the back of the tongue without gagging were screened using the same

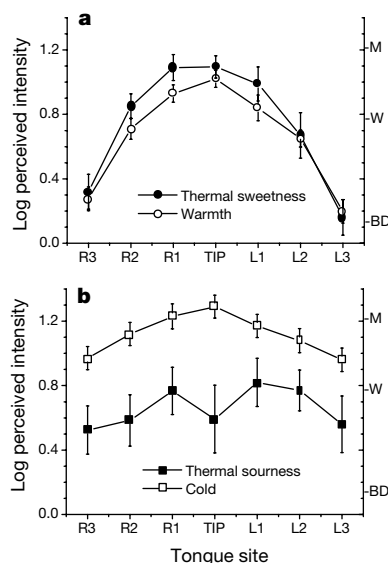


Figure 4 Thermal sweetness and warmth (a), and thermal sourness and cold (b) as a function of test site along the anterior edge of the tongue. The stimulus for warming was an increase from 20 to 35 °C; for cooling it was a decrease from 35 to 15 °C. R1–R3 and L1–L3 refer to contiguous test sites (~8 mm wide) along the right and left sides of the tongue, respectively. The data are from those subjects whose average rating of sweetness (a, $n = 14$) or sourness (b, $n = 13$) exceeded 'barely detectable' at one or more sites.

thermal conditions as on the front of the tongue. Ten of these subjects reported thermal taste, but systematic tests showed that its characteristics were different than in the fungiform region: warming (Fig. 5a) failed to produce even barely detectable sweetness whereas cooling (Fig. 5b and c) elicited thermal bitterness (T_{BI}) as well as T_{SO} (but not T_{SA}), with half the subjects reporting bitterness and half reporting sourness. The appearance of T_{BI} was particularly striking in that subjects who experienced it had reported sourness under the same conditions in the fungiform region. These differences appear unrelated to possible regional differences in sensitivity to chemical taste, as intensity ratings obtained for all four taste solutions on the circumvallate papillae were comparable to those previously obtained on the front of the tongue. In particular, 0.5 M sucrose evoked approximately moderate sweetness on circumvallate papillae, just as it did on the tongue tip. Consistent with the close association between T_{SW} and warmth on the front of the tongue, warmth sensitivity was poor on circumvallate papillae, whereas cooling was rated similarly in both regions.

The most straightforward explanation of thermal taste is that temperature-sensitive neurons in the human chorda tympani² and glossopharyngeal nerves encode taste rather than temperature. Electrophysiological studies in animals have shown that cold-sensitive gustatory neurons tend to be sensitive to acids (sour) or salts, and that warm-sensitive neurons tend to be sensitive to sucrose (sweetness) or bitter substances^{3,4}. Because sensitivity to sweet and bitter chemicals depends upon G-protein-coupled receptors (GPCRs)^{12,13} and sensitivity to salts and acids depends upon specialized Na^+ and H^+ ion channels¹⁴, it is possible some GPCR cascades can be triggered by warming and that certain Na^+ and H^+ channels⁵ can be gated by cooling. If so, the large individual and spatial differences in thermal taste we observed may reflect variations

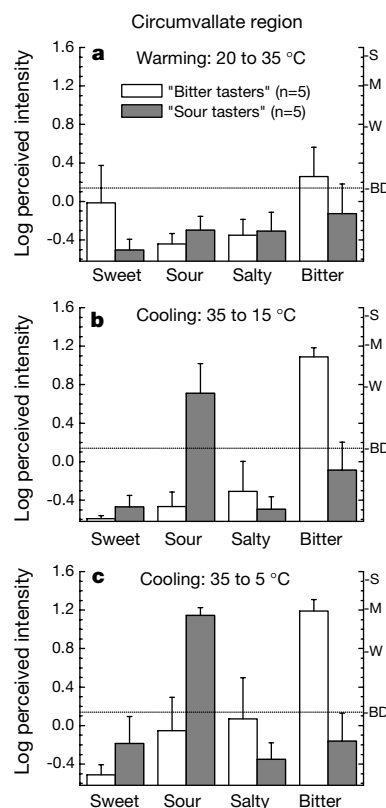


Figure 5 Thermal taste ratings from the circumvallate region of the tongue. Data are shown for 10 subjects in three temperature conditions: a, warming from 20 to 35 °C; b, cooling from 35 to 15 °C; c, cooling from 35 to 5 °C. Subjects were grouped according to the dominant taste quality they reported (bitter or sour) during cooling.

in the incidence and distribution of neurons whose chemosensory mechanisms are temperature-sensitive. However, the absence of significant bitterness during warming and the reports of bitterness during cooling on circumvallate papillae raise the possibility that thermal sensitivity in some gustatory neurons may arise from cellular processes that are unrelated to chemosensory transduction.

We note that thermal taste was nearly discovered 35 years ago by von Békésy. In a well-known but controversial paper, von Békésy¹⁵ reported that taste and thermal stimuli (heated or cooled water) presented to opposite sides of the tongue merged into a single sensation when warm water was paired with sucrose or quinine, or when cold water was paired with citric acid or NaCl. This observation led him to propose the 'Duplexity Theory of Taste'¹⁵, in which he posited that "warm and cold stimuli act similarly to the four primary taste stimuli..." Our results now suggest that von Békésy's subjects may have reported a single sensation in the middle of the tongue when bilateral thermal and chemical stimuli evoked the same taste quality. □

Methods

Thermal taste screening procedure

The incidence of thermal taste was tested in naive subjects (8 males and 16 females, most of whom were students at Yale University) using three temperature conditions that pilot tests had shown were capable of producing sweetness, sourness and saltiness, respectively: warming from 20 to 35 °C, cooling from 35 to 15 °C, and cooling from 35 to 5 °C. Temperature was varied at approximately $\pm 1.5^\circ\text{C s}^{-1}$ using an 8 mm $\times$ 8 mm computer-controlled Peltier thermode with thermocouple feedback. The thermode was affixed to a pencil-sized water-circulated heat sink and covered with plastic wrap for hygienic purposes. On each trial the thermode was set to the starting temperature, and with guidance from the experimenter and the aid of a mirror, subjects used the heat sink as a handle to position the thermode against the tongue. Heating or cooling began as soon as the temperature at the tongue-thermode interface stabilized at the starting temperature (5–10 s). Subjects were told to attend to the temperature change and to report if they perceived any other sensations, including tastes (defined as sweetness, sourness, saltiness or bitterness); they were assured that not everyone perceived such sensations, and that the purpose of the study was to discover how often and under what conditions they might appear. Stimulation began on the tongue tip and proceeded stepwise along the edge of the tongue to a distance ~ 5 cm caudal to the tip. Both sides of the tongue were tested, and each temperature condition was applied twice to each test site. When tastes were detected subjects reported their intensities verbally using a scale from 1 to 10. These ratings served to locate 'best' sites for thermal taste that were later tested more systematically.

Thermal testing on the tongue tip

The thermode was used to warm or cool the tongue tip over a series of temperature steps (ΔT s) that increased from 20 °C in steps (°C) of +5, +10, +15 and +20, or decreased from 35 °C in steps (°C) of –10, –15, –20, –25 and –30. Subjects rated the intensity of taste (sweetness, sourness, saltiness, bitterness) and thermal sensations (warmth, cold) using the labelled magnitude scale (LMS)¹⁶, a continuous scale of sensation intensity bounded by 'no sensation' and 'strongest imaginable oral sensation'. The LMS was displayed on a computer monitor and subjects made their ratings using a mouse. Instructions were given to "attend now" as soon as heating began on warming trials and as soon as the target temperature was reached on cooling trials. Different instructions were used for heating and cooling because pilot tests had shown that sweetness occurred only while temperature rose, whereas sourness and saltiness persisted at steady temperatures. Because thermal taste was always accompanied by temperature sensations, taste and temperature ratings were obtained separately to help subjects make independent judgments. Each condition was presented twice in pseudo-random sequence.

Testing on 'best' thermal taste sites

18 subjects (one of the original 19 left the study between experiments) rated thermal tastes and temperature sensations in the same manner as on the tongue tip, except temperature was varied only as follows: from 20 to 35 °C to assess T_{SW} , from 35 to 15 °C to assess T_{SO} , and from 35 to 5 °C to assess T_{SA} . Chemical taste was assessed in a separate session on the same sites using four aqueous taste solutions (0.5 M sucrose, 0.1 M citric acid, 0.5 M NaCl and 0.01 M QHCl) found in pilot tests to produce approximately 'moderate' sweetness, saltiness, sourness or bitterness, respectively, when applied to small areas of the tongue. The experimenter used cotton-tipped applicators to carefully swab these solutions onto T_{SW} and T_{SO} 'best' sites for 3 s. Subjects used the LMS to rate intensity and rinsed between trials with distilled H₂O. Two replicates were obtained for each thermal and chemical condition.

Stepwise spatial testing

Measurements of T_{SW} and T_{SO} on the edge of the tongue were made on another group of 15 subjects (12 females and 3 males, screened as before from a sample of 22 females and 8

males). Seven sites were tested: the tongue tip and three contiguous locations on either side of the tip. Stimulation began at the tip and stepped approximately one width of the thermode (8 mm) at a time, first along one side of the tongue and then the other. At each site the thermode was warmed from 20 to 35 °C or cooled from 35 to 15 °C, with cooling and warming trials blocked. Two replicates were obtained for each temperature condition at each site.

Received 4 November; accepted 8 December 1999.

1. Zotterman, Y. Action potentials in the glossopharyngeal nerve and in the chorda tympani. *Skand. Arch. Physiol.* **72**, 73–77 (1935).
2. Oakley, B. Taste responses of human chorda tympani nerve. *Chem. Senses* **10**, 469–481 (1985).
3. Ogawa, H., Sato, M. & Yamashita, S. Multiple sensitivity of chorda tympani fibres of the rat and hamster to gustatory and thermal stimuli. *J. Physiol.* **199**, 223–240 (1968).
4. Sato, M., Ogawa, H. & Yamashita, S. Response properties of macaque monkey chorda tympani fibers. *J. Gen. Physiol.* **66**, 781–810 (1975).
5. Nakamura, M. & Kurihara, K. Temperature dependence of amiloride-sensitive and -insensitive components of rat taste nerve response to NaCl. *Brain Res.* **444**, 159–164 (1988).
6. Travers, S. P. & Smith, D. V. Responsiveness of neurons in the hamster parabrachial nuclei to taste mixtures. *J. Gen. Physiol.* **84**, 221–250 (1984).
7. McBurney, D. H., Collings, V. B. & Glanz, L. M. Temperature dependence of human taste response. *Physiol. Behav.* **11**, 89–94 (1973).
8. Bartoshuk, L. M., Rennert, K., Rodin, H. & Stevens, J. C. Effects of temperature on the perceived sweetness of sucrose. *Physiol. Behav.* **28**, 905–910 (1982).
9. Green, B. G. & Frankmann, S. P. The effect of cooling the tongue on the perceived intensity of taste. *Chem. Senses* **12**, 609–619 (1987).
10. Whitehead, M. C., Ganchrow, J. R., Ganchrow, D. & Yao, B. Organization of geniculate and trigeminal ganglion cells innervating single fungiform taste papillae: a study with tetramethylrhodamine dextran amine labeling. *Neuroscience* **93**, 931–941 (1999).
11. Whitehead, M. C. & Kachele, D. L. Development of fungiform papillae, taste buds, and their innervation in hamster. *J. Comp. Neurol.* **340**, 515–530 (1994).
12. Wong, T. G., Gannon, K. S. & Margolskee, R. F. Transduction of bitter and sweet taste by gustducin. *Nature* **381**, 796–800 (1996).
13. Hoon, M. A. et al. Putative mammalian taste receptors: A class of taste-specific GPCRs with distinct topographic selectivity. *Cell* **96**, 541–551 (1999).
14. Lindemann, B. Taste reception. *Physiol. Rev.* **76**, 718–766 (1996).
15. von Békésy, G. Duplexity theory of taste. *Science* **145**, 834–835 (1964).
16. Green, B. G., Shaffer, G. S. & Gilmore, M. M. Derivation and evaluation of a semantic scale of oral sensation magnitude with apparent ratio properties. *Chem. Senses* **18**, 683–702 (1993).

Acknowledgements

We thank A. Hoffmann for collecting and analysing some of the data, and R. Rascati, F. Strumpf and M. Fritz for technical assistance.

Correspondence and requests for materials should be addressed to B.G.G. at the John B. Pierce Laboratory (e-mail: green@jbpierce.org).

Postsaccadic visual references generate presaccadic compression of space

Markus Lappe, Holger Awater & Bart Krekelberg

Department of Zoology and Neurobiology, Ruhr-University Bochum, 44780 Bochum, Germany

With every rapid gaze shift (saccade), our eyes experience a different view of the world. Stable perception of visual space requires that points in the new image are associated with corresponding points in the previous image. The brain may use an extraretinal eye position signal to compensate for gaze changes^{1,2}, or, alternatively, exploit the image contents to determine associated locations^{3,4}. Support for a uniform extraretinal signal comes from findings that the apparent position of objects briefly flashed around the time of a saccade is often shifted in the direction of the saccade^{5–9}. This view is challenged, however, by observations that the magnitude^{4,10} and direction¹¹ of the displacement varies across the visual field. Led by the observation that non-uniform displacements typically occurred in studies conducted in slightly illuminated rooms^{4,7,10–13}, here we determine

the dependence of perisaccadic mislocalization on the availability of visual spatial references at various times around a saccade. We find that presaccadic compression¹¹ occurs only if visual references are available immediately after, rather than before or during, the saccade. Our findings indicate that the visual processes of transsaccadic spatial localization use mainly postsaccadic visual information.

We asked five observers to locate a briefly flashed (8-ms) luminous bar on a projection screen in a dark room while they were making saccades to the right. The bar was flashed at variable times before or after saccade onset. It appeared randomly at one of four locations around the saccade goal. Subjects reported the perceived location of the bar with a mouse pointer that appeared 500 ms after the saccade. A ruler that was projected on the screen provided visual references. The ruler could be switched on or off at various times during a trial. This allowed us to choose the times at which visual references were available. Figure 1 shows responses obtained from a subject when the ruler was available. Locations between the fixation point and the saccade goal are mislocalized in the direction of the saccade. Locations beyond the saccade goal are mislocalized against the direction of the saccade. The mislocalization starts about 70 ms before the saccade and continues during the saccade, until about 70 ms after saccade onset. We defined two index measures to compare the mislocalizations across subjects and conditions. The shift index describes the overall perceived shift in the direction of the saccade. It is defined as the mean over the four mean apparent positions of the bar. The compression index describes the strength of the compression. It is defined as the standard deviation of the four mean apparent positions. Both indices are normalized to their respective average values 100 ms before and after the saccade. Because the individual index curves were similar across subjects we averaged their data. Figure 2 shows the perisaccadic time course of these two measures when the ruler was present (blue curves). There is strong compression. In addition, there is also a shift in the mean apparent position. When we removed the ruler from the screen (that is, when no visual references were available) the compression was much weaker (Fig. 2, red curves). The shift is slightly larger in this case. This clearly shows that the pattern of perisaccadic mislocalizations is influenced by visual references. Without visual references, mislocalization is more uniform in direction and magnitude. This presumably reflects a unary extraretinal signal which is the only available information about the change of gaze direction in this case. When visual references are present, the metric of the mislocalization changes. All perceived positions move closer together and cluster around the saccade goal.

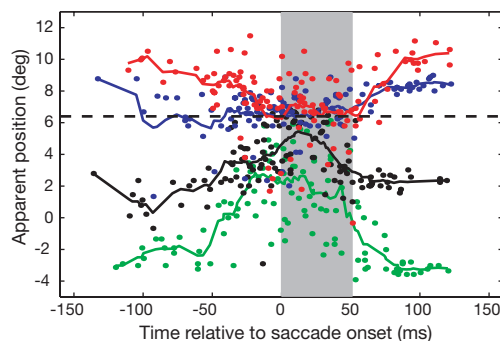


Figure 1 Perceived location of a flashed bar at position -2.6° (green), $+2.6^\circ$ (black), $+10^\circ$ (blue) and $+13^\circ$ (red) as a function of time relative to the onset of a saccade from -6.4° to $+6.4^\circ$. Each point is a single measurement. Lines are running averages through the data, obtained with a Gaussian filter of 33-ms standard deviation. Dashed line indicates the saccade goal. Around the time of the saccade, perceived locations are strongly biased towards the saccade goal.

What aspect of the visual references provided by the ruler induces this change? The ruler may serve as a stable transsaccadic reference frame that provides matching visual locations before and after the saccade. If this were true, one would predict less compression if the ruler was absent either before or after the saccade. It could also be that the ruler directly provides a reference frame for the flash. In this case one would expect less compression if the ruler was absent at the time of the flash. Finally, it might be that the visual motion of the ruler during the saccade is important to induce the compression. In this case one would expect less compression if the ruler was absent during the saccade, but not before or after it. To investigate these possibilities we ran a series of further experiments in which the ruler was switched on or off at different times during the trial. In each case we determined the mean presaccadic compression by averaging the compression index values of each subject between 50 and 0 ms before the saccade, and then took the mean across subjects. The results are expressed as percentages with 100% compression corresponding to the percept where all flashes are seen at the same place (Fig. 3).

We first tested whether the ruler acts as a stable transsaccadic reference frame. We presented the ruler selectively before or after the flash of the bar. In the pre-flash condition, the ruler was on at the beginning of the trial, stayed on until the presentation of the flashed bar and then went off with the bar. In the post-flash condition, the ruler was off initially and was switched on when the flashed bar appeared. We found very little compression in the pre-flash condition but strong compression in the post-flash condition (Fig. 3). We conclude that the compression does not rely on a stable transsaccadic reference. Rather, it must be induced by visual references present after the flash.

In two further experiments we tested whether the presence of the ruler during the saccade or immediately after the flash is important. In the gap condition, the ruler was on before and after the saccade but was switched off for 250 ms starting from the flash of the bar. Some compression occurred but it was not significantly different from that found when the ruler was absent either throughout the trial or after the flash (Fig. 3). In the post-saccade condition the

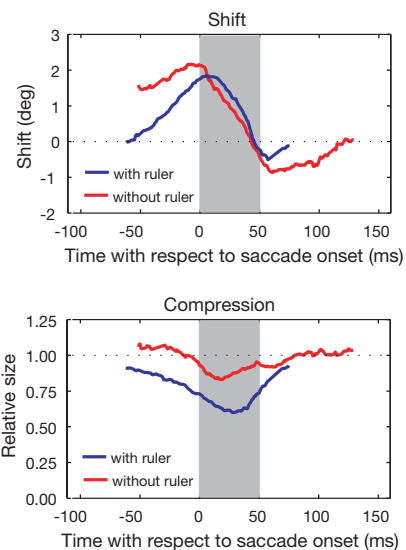


Figure 2 Perceived shift and compression as a function of time in the presence (blue) and absence (red) of visual references. Perceived shift is the mean of the four mean apparent positions of the bar relative to the mean apparent positions more than 100 ms before or after the saccade. Perceived compression ('relative size') is the standard deviation of the four mean apparent positions relative to the respective value more than 100 ms before or after the saccade. A value of 1 indicates no compression, a value of 0 would occur if all four positions appeared in a single place. Shift and compression were calculated from the individual localization curves of five subjects and then averaged.

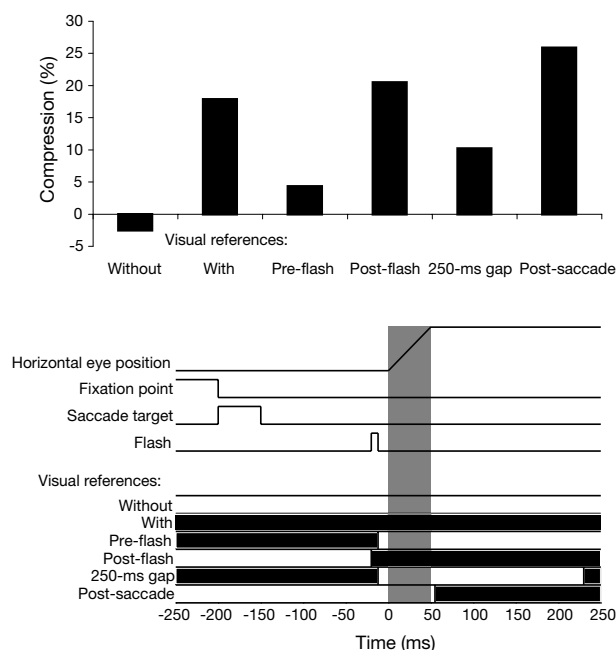


Figure 3 Magnitude of presaccadic compression for different presentation times of visual references. Top, height of bar is proportional to the mean of the compression measurements from -50 to 0 ms before the saccade, averaged across subjects. Compression in the with, post-flash and post-saccadic conditions is significantly different from that in the without and pre-flash conditions (analysis of variance and Student–Newman–Keuls post-hoc testing, $P < 0.05$). Bottom, timing of events in the different conditions. Black areas show the times when the ruler was present. The first two conditions (without/with visual references) are the same as in Fig. 2. In the pre-flash condition, visual references were only available before the flash. In the post-flash condition, visual references were initially absent but appeared with the flash and stayed thereafter. In the gap condition, visual references were turned off for 250 ms starting from the flash of the bar. In the post-saccadic condition, visual references appeared only after the saccade.

ruler was initially absent and switched on immediately after the saccade. We used the eye movement to trigger the presentation of the ruler. The stimulation program detected the start of the saccade and switched on the ruler 50 ms later, at the time when the saccade had just finished. This condition gave strong compression (Fig. 3). We conclude that visual references or visual motion during the saccade are not important. Rather, the presence of visual references immediately after the saccade elicits compression.

These results reconcile the seemingly disparate findings of refs 9, 10. In ref. 9, where the compression was described, continuous visual references were available, such as the visible frame of the monitor and the constantly present fixation and saccade targets. In ref. 10, which reported only shifts in the direction of the saccade, all potential visual references (the fixation point, saccade goal and comparison targets) were extinguished before the saccade.

Why then does the presence of postsaccadic visual references lead to a compression of presaccadic space, which is not seen in the absence of visual references? First, it is important to realize that an observer faces different problems in these two conditions¹³. To refer the presaccadic retinal signal to the appropriate world coordinates in the absence of visual references, the visual system can only subtract an eye-position signal. No sources of information beyond this signal are available, so any inaccuracies in this signal will be translated into mislocalizations^{3,5–9}. In the presence of postsaccadic visual references, on the other hand, the task changes from egocentric localization to a relative position judgement in a retinal frame of reference, that is, with respect to other visible targets. We believe that different mechanisms are involved in these tasks¹⁴. Second, around the time of the saccade, mechanisms underlying

perceptual stability are operating. Perceptual stability relies strongly on the visual information present immediately after the saccade finishes¹⁵. At that time, the postsaccadic location of the saccade target is determined from the visual scene¹⁶. In abstract terms, the internal presaccadic coordinate system is being remapped to comply with the postsaccadic eye position. Owing to the latency of visual processing and the anticipatory nature of remapping¹⁷, a presaccadic stimulus could be interpreted in a visual coordinate system that is being remapped, or in one that has already reached its postsaccadic state. In either case a mislocalization will result. In this interpretation, the presaccadic compression of space is a signature of the fact that the visual remapping process is not instantaneous or uniform over the retina and that it requires early postsaccadic image information. □

Methods

Subjects sat in a dark room (luminance $< 0.1 \text{ cd m}^{-2}$) in front of a large projection screen. Visual stimuli were generated by a computer and presented on the projection screen by a video projector with a frame rate of 120 Hz. Each trial started with the appearance of a fixation dot 6.4° left of the screen centre followed after 1,700–1,870 ms by the appearance of a dot marking the saccade goal 6.4° right of the centre. The fixation point and saccade goal were extinguished 50 ms later (Fig. 3, lower panel). The subject was required to make a saccade from the fixation point to the saccade goal. Because saccadic latencies were between 110 and 190 ms, neither the fixation point nor the saccade target were visible at the time of the saccade. At a random time within 250 ms after the appearance of the saccade goal, a vertical bar ($0.5^\circ \times 90^\circ$, mean luminance 20 cd m^{-2}) was flashed at one of four possible positions ($-2.6, 2.6, 10$ or 13.6°) for one video frame (8 ms). About 500 ms after the saccade, a mouse pointer appeared. The subject moved the mouse pointer to the apparent horizontal location of the bar and pressed a button. This location was recorded along with the time of the flash. Then the next trial started.

Visual references were provided by a horizontal ruler displayed on the screen. The ruler was a horizontal white line (luminance 20 cd m^{-2}) with short vertical lines at 12.8° intervals, each labelled with a number. One of these marks fell on the fixation point, another on the saccade goal. The ruler extended over the entire width of the screen. In the first experiment, the ruler was continuously visible. In the second experiment it was completely absent. In the remaining four experiments the ruler was switched on or off at different times during the trial. The lower panel of Fig. 3 shows the timing of events in the different conditions. All subjects performed the complete set of experiments. Between 170 and 300 responses were collected per subject and condition.

Eye movements were measured with an Ober 2 infrared eye tracker at a sample rate of 200 Hz. Saccade initiation time was determined by a velocity criterion with a threshold of 10% of the maximum speed during the saccade. Every saccade was visually checked by the experimenter for appropriate direction, amplitude and timing. Trials in which the saccade did not meet the requirements of the task were discarded. The last experiment used additional electro-oculography to trigger the appearance of the ruler with the eye movement.

Received 11 October; accepted 15 December 1999.

1. von Helmholtz, H. *Handbuch der Physiologischen Optik* (Leopold Voss, Hamburg, 1896).
2. Von Holst, E. & Mittelstaedt, H. Das Reafferenzprinzip (Wechselwirkung zwischen Zentralnervensystem und Peripherie). *Naturwissenschaften* **37**, 464–476 (1950).
3. MacKay, D. M. Mislocation of test stimuli during saccadic image displacement. *Nature* **227**, 731–733 (1970).
4. O'Regan, J. K. Retinal versus extraretinal influences in flash localization during saccadic eye movements in the presence of a visual background. *Percept. Psychophys.* **36**, 1–14 (1984).
5. Matin, L. & Pearce, D. G. Visual perception of direction for stimuli during voluntary saccadic eye movements. *Science* **148**, 1485–1488 (1965).
6. Honda, H. Perceptual localization of visual stimuli flashed during saccades. *Percept. Psychophys.* **45**, 162–174 (1989).
7. Dassonville, P., Schlag, J. & Schlag-Rey, M. The use of egocentric and exocentric location cues in saccadic programming. *Vision Res.* **35**, 2191–2199 (1995).
8. Bockisch, C. & Miller, J. M. Different motor systems use similar damped extraretinal eye position information. *Vision Res.* **39**, 1025–1038 (1999).
9. Cai, R. H., Pouget, A., Schlag-Rey, M. & Schlag, J. Perceived geometrical relationships affected by eye-movement signals. *Nature* **386**, 601–604 (1997).
10. Bischof, N. & Kramer, E. Untersuchungen und Überlegungen zur Richtungswahrnehmung bei willkürlichen sakkadischen Augenbewegungen. *Psychologische Forschung* **32**, 185–218 (1968).
11. Ross, J., Morrone, M. C. & Burr, D. C. Compression of visual space before saccades. *Nature* **386**, 598–601 (1997).
12. Honda, H. Saccade-contingent displacement of apparent position of visual stimuli flashed on a dimly illuminated structured background. *Vision Res.* **33**, 709–716 (1993).
13. Miller, J. M. & Bockisch, C. Where are the things we see? *Nature* **386**, 550–551 (1997).
14. Krekelberg, B. & Lappe, M. A model of the perceived relative positions of moving objects based upon a slow averaging process. *Vision Res.* **40**, 201–215 (2000).
15. Deubel, H., Schneider, W. X. & Bridgeman, B. Postsaccadic target blanking prevents saccadic suppression of image displacement. *Vision Res.* **36**, 985–996 (1996).
16. Deubel, H., Bridgeman, B. & Schneider, W. X. Immediate post-saccadic information mediates space constancy. *Vision Res.* **38**, 3147–3159 (1998).
17. Duhamel, J.-R., Colby, C. L. & Goldberg, M. E. The updating of the representation of visual space in parietal cortex by intended eye movements. *Science* **255**, 90–92 (1992).

Acknowledgements

We thank L. Lünenburger and M. Klar for their help. Financial support from the Human Frontier Science Program is gratefully acknowledged.

Correspondence and requests for materials should be addressed to M.L. (e-mail: lappe@neurobiologie.ruhr-uni-bochum.de).

A neurofibromatosis-1-regulated pathway is required for learning in *Drosophila*

Hui-Fu Guo*, Jiayuan Tong*, Frances Hannan, Lin Luo & Yi Zhong

Cold Spring Harbor Laboratory, P.O. Box 100, Cold Spring Harbor, New York 11724, USA

* These authors contributed equally to this work

The tumour-suppressor gene *Neurofibromatosis 1* (*Nf1*) encodes a Ras-specific GTPase activating protein (Ras-GAP)^{1–5}. In addition to being involved in tumour formation^{6,7}, *Nf1* has been reported to cause learning defects in humans^{8–10} and *Nf1* knockout mice¹¹. However, it remains to be determined whether the observed learning defect is secondary to abnormal development. The *Drosophila* NF1 protein is highly conserved, showing 60% identity of its 2,803 amino acids with human NF1 (ref. 12). Previous studies have suggested that *Drosophila* NF1 acts not only as a Ras-GAP but also as a possible regulator of the cAMP pathway that involves the *rutabaga* (*rut*)-encoded adenylyl cyclase¹³. Because *rut* was isolated as a learning and short-term memory mutant^{14,15}, we have pursued the hypothesis that NF1 may affect learning through its control of the Rut-adenylyl cyclase/cAMP pathway. Here we show that NF1 affects learning and short-term memory independently of its developmental effects. We show that G-protein-activated adenylyl cyclase activity consists of NF1-independent and NF1-dependent components, and that the mechanism of the NF1-dependent activation of the Rut-adenylyl cyclase pathway is essential for mediating *Drosophila* learning and memory.

We examined olfactory associative learning of adult fruit flies by using a well-defined Pavlovian procedure^{16–19}. Significant decrements in olfactory learning performance were shown for two independently isolated *NF1* null alleles¹², *NF1^{P1}* and *NF1^{P2}*, as compared with K33, the parental line for *NF1* mutants with a P-element inserted nearby the *NF1* locus¹² (Table 1, Fig. 1a). Olfactory avoidance and electric-shock reactivity²⁰, two sensorimotor activities necessary for performing the learning task, were similar in the mutant and control K33 flies (Table 1). To consider the potential

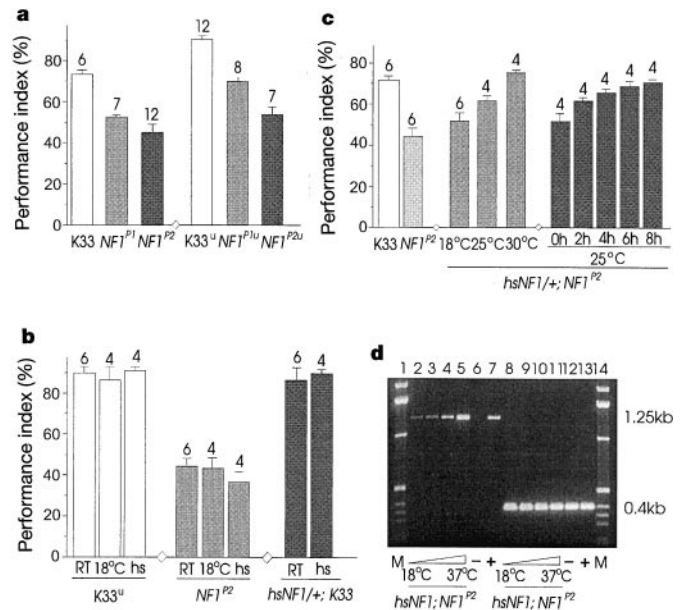


Figure 1 Rescue of the *NF1* learning defect by inducible expression of the normal *NF1* transgene. **a**, *NF1* learning defects observed in both the original and outcrossed isogenic (marked by u in superscript) genetic background. K33 is the parental line of *NF1* mutants. **b**, No effect on learning scores for the heat-shock treatment in controls, including overexpression of the *NF1* transgene in the control background. **c**, Rescue of the learning defect by induced expression of the *NF1* transgene. In the first group, the flies were moved from 18°C to 25 or 30°C for 2 h before the learning test ($P < 0.05$, Tukey Kramer Honestly Significant Difference). In the second group, flies were shifted from 18°C to 25°C for 0, 2, 4, 6 or 8 h, respectively (significant for 2 h, $P < 0.05$). The number of assays for each group are indicated above each error bar. **d**, Semi-quantitative RT-PCR showing induced expression of the *hsNF1* transgene. Lanes 1 and 14, 1-kb DNA ladder (M) (Gibco BRL). Lanes 2–7, RT-PCR using *NF1*-specific primers with cDNA prepared from *hsNF1*; *NF1^{P2}* flies grown at 18, 25 and 30°C or given daily 1 h heat shock at 37°C, or from *NF1^{P1}* mutant (–) flies or K33 wild-type (+) flies grown at 18°C. Lanes 8–13, control RT-PCR from the same cDNA using ribosomal protein *rp49*-specific primers. Three separate mRNA isolations showed the same pattern of increased expression of the *hsNF1* transgene at increased temperature.

effects of genetic background on behaviour²⁰, we outcrossed *NF1* mutants and K33 with an isogenic line *w¹¹¹⁸* (*isoCJ1*)²¹. Again, learning scores of *NF1* mutants were significantly reduced (Table 1, Fig. 1a), whereas the parameters of sensorimotor activities were not statistically different from the control with a similar genetic background (Table 1). Even though learning scores and some scores for shock reactivity and odour avoidance are significantly different for K33 in different genetic backgrounds, these behavioural parameters also vary accordingly in *NF1* mutants (Table 1). These results indicate that *NF1* is a learning mutant.

Table 1 Performance index for olfactory learning, shock reactivity and odour avoidance

Genotypes	Learning (n)	Odour avoidance					
		Shock reactivity		BA dilution		MCH dilution	
		60 V	20 V	4%	0.4%	Undiluted	10%
K33	73 ± 2 (6)	72 ± 6	25 ± 9	78 ± 5	28 ± 10	73 ± 7	40 ± 9
<i>NF1^{P1}</i>	53 ± 1 (7)*	77 ± 4	30 ± 5	80 ± 3	25 ± 4	66 ± 8	36 ± 6
<i>NF1^{P2}</i>	45 ± 4 (12)*	76 ± 3	26 ± 5	71 ± 4	19 ± 6	67 ± 4	29 ± 7
<i>hsNF1/+; NF1^{P2}</i>	75 ± 2 (4)	65 ± 3	23 ± 6	79 ± 6	32 ± 11	83 ± 4	34 ± 7
K33 ^u	90 ± 1 (12)	89 ± 2	61 ± 6	92 ± 1	41 ± 8	77 ± 5	63 ± 5
<i>NF1^{P1}u</i>	70 ± 2 (8)	80 ± 5	56 ± 8	85 ± 5	36 ± 9	84 ± 6	59 ± 8
<i>NF1^{P2}u</i>	54 ± 4 (7)*	83 ± 3	62 ± 7	93 ± 2	41 ± 7	77 ± 4	58 ± 5

K33, *NF1^{P1}*, *NF1^{P2}* and *hsNF1/+; NF1^{P2}* have a similar genetic background, whereas K33^u, *NF1^{P1}u* and *NF1^{P2}u* have a different background (see Methods). All scores are expressed as $\bar{x} \pm \text{s.e.m.}$ For learning, the number (n) of assays are indicated in parentheses. For all shock reactivity and odour avoidance assays, $n = 8$.

* Statistically different from control. No statistical difference at the level of $\alpha = 0.05$ is detected among all the sensorimotor activities. Learning defect is significant at $\alpha \leq 0.001$. Comparison is made between mutants and controls with a similar genetic background using Tukey–Kramer HSD test within the Macintosh software package JMP3.1 (SAS institute, Inc., Cary, North Carolina, USA).

This conclusion is further supported by the observation that the learning defect was rescued by induced expression of the *NF1* transgene (see below) without changing sensorimotor activity significantly (Table 1).

We then examined the effect of heat-shock-induced expression of the *NF1* transgene to determine whether the learning defect is caused by an adult requirement for NF1, or whether it is a secondary consequence of developmental abnormalities, such as the small body size of *NF1* mutants¹². Heat-shock treatment of *hsNF1* transgenic flies leads to expression of the NF1 protein¹². Such treatment did not affect learning scores in *NF1* mutants, control flies or *hsNF1*; *K33* (Fig. 1b); however, learning scores were improved when mutant flies carrying the *NF1* transgene were heat shocked.

Heterozygous transgenic *NF1* (*hsNF1/+*; *NF1*^{P2}) flies were raised at room temperature (20–24 °C). These flies showed a learning score PI of 63 ± 3 ($n = 5$), indicating a partial rescue because of leaky expression (see Fig. 1d). To minimize the leaky expression of the heat-shock promoter-controlled *NF1* transgene, flies were shifted to 18 °C overnight before the test. This reduced the learning score significantly to 52 ± 4 (Fig. 1c). Learning scores of transgenic flies (*hsNF1/+*; *NF1*^{P2}) were improved to a better extent when flies were treated at 30 °C as compared with 25 °C for two hours (Fig. 1c), or when the transgenic flies were subjected to 25 °C for successively longer times (Fig. 1c). Presumably, more NF1 was expressed with higher temperatures or for longer times of treatment. This is supported by data from polymerase chain reaction with reverse transcriptase (RT-PCR). Higher temperature treatment led to

accumulation of more messenger RNA transcribed from the *hsNF1* transgene (Fig. 1d). Thus, the level of performance improvement may be proportional to the amount of NF1 expressed.

A single heat-shock treatment during larval stages did not change the smaller body size of *NF1* mutants, as quantified by measuring the length of pupal cases. *hsNF1/+*; *NF1*^{P2} pupal cases (2.73 ± 0.14 ; $n = 66$) were indistinguishable after treatment from *NF1*^{P2} (2.69 ± 0.2 ; $n = 56$), but were smaller than those of control *K33* (3.2 ± 0.13 ; $n = 70$). Repetitive or continuous heat-shock treatments, however, rescue the developmental phenotype¹². Thus, the learning defect can be rescued by acute expression of the *NF1* transgene during adulthood, but the developmental defect requires repetitive or continuous heat-shock treatment during development. This suggests that NF1 is essential for the learning process.

To test whether the NF1-dependent learning defect involves the cAMP pathway, we compared learning scores of *NF1*^{P2} and *rut*¹ single-mutant, and *rut*¹; *NF1*^{P2} double-mutant flies. The learning scores of all three mutant genotypes were very similar (Fig. 2a). The learning score of another double mutant, *dunce* (*dnc*); *rut*¹, is reduced when compared with either single mutants¹⁶, which indicates that the two mutations exert additive effects on learning even though both gene products are involved in the cAMP cascade (Rut-adenylyl cyclase (AC) for synthesizing cAMP^{14,22} and Dnc-phosphodiesterase for degrading cAMP^{23,24}). Therefore, the absence of any further reduction of learning in the double mutant *rut*¹; *NF1*^{P2} suggests that both gene products function closely in the cAMP pathway.

This idea is supported by studies of *NF1* mutant flies carrying a transgene encoding a mutant catalytic subunit of cAMP-dependent protein kinase (PKA*), which is constitutively active²⁵. Sustained expression of this PKA subunit rescues the small body size phenotype of *NF1* mutants¹². Heat-shock induction of the constitutively active PKA should, in principle, bypass the requirement for the Rut-AC and all other molecules upstream of normal PKA activation. The *hsp70-PKA** transgene completely rescued the learning defect of *NF1*^{P1} when the flies were raised at room temperature (Fig. 2b). *NF1*^{P2} mutants were partially rescued by the transgene at room temperature, but showed complete rescue with heat shock (37 °C,

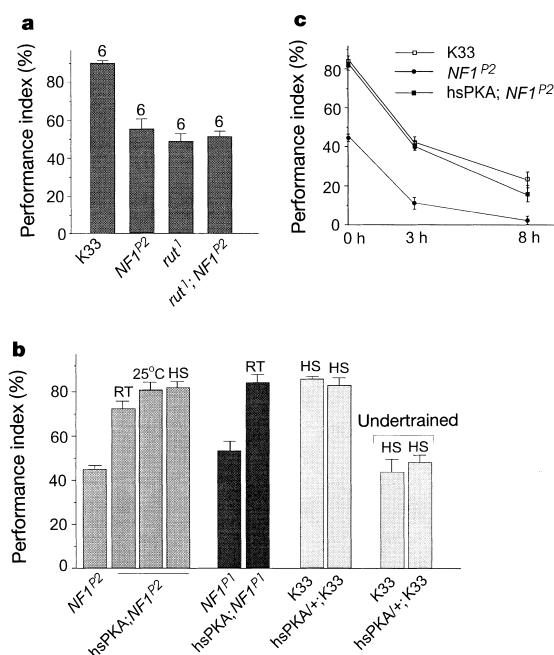


Figure 2 Effects of the cAMP pathway on NF1-dependent learning and memory. **a**, Effect of the *rut* mutation. Learning defects were similar among *rut*, *NF1* and *rut*¹; *NF1*^{P2} mutants. **b**, Rescue of *NF1* learning defects by induced expression of a constitutively active catalytic subunit of cAMP-dependent protein kinase (PKA*). Learning scores were rescued or partially rescued in *hsp70-PKA**; *NF1*^{P1} and *hsp70-PKA**; *NF1*^{P2} flies, even when raised at room temperature (RT). For undertraining, flies after heatshock (HS) were trained with 3 repeats of electric shock instead of 12 to avoid any ceiling effect on learning scores. The learning scores were reduced in *K33* and *hsPKA*⁺; *K33* in parallel. Heat shock was at 37 °C for 30 min and training started after a 3-h rest. Flies were shifted to 25 °C overnight before training (25 °C). From left to right, $n = 6, 6, 6, 5, 4, 6, 6, 4, 4$. **c**, Rescue of short-term memory. Retention at 3 and 8 h after training was also disrupted in *NF1* mutants. This was rescued by induced expression of the PKA* subunit. Flies were raised at room temperature and heat shocked for 30 min at 37 °C and then rested for at least 3 h before training. $n = 6-12$.

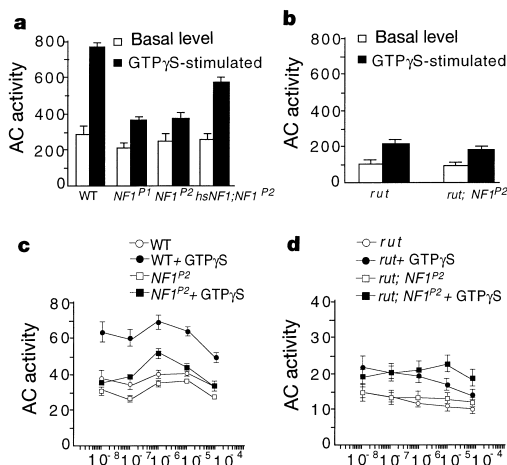


Figure 3 Biochemical assay of the effects of NF1 on AC activity. **a**, Reduction of GTPγS-stimulated AC activity in brain tissue of *NF1* mutants. Each data point (mean $\pm$ s.e.m.) is the average of four independent experiments. To be comparable, all flies were subjected to heat shock (for 2 h at 35 °C and 1 h rest at room temperature). **b**, Diminished *NF1* effect on AC activity in the *rut* mutant background. Flies were also subjected to the same heat shock as in **a** and brain tissues were used. Data points are the average of three independent experiments. **c**, Effects of NF1 on Ca^{2+} dependence of AC activity in abdominal tissues. The average of eight independent experiments is shown. **d**, No significant *NF1* effect on abdominal AC activity in the *rut* mutant background. The average of six independent experiments is shown.

30 min), or with a shift to 25°C overnight before being tested (Fig. 2b). In addition, *NF1* mutations also caused a short-term memory defect (3- and 8-h retention, Fig. 2c; *NF1^{P1}* not shown), which was also fully rescued by heat-shock induction of PKA*. To determine whether expression of *hsp70-PKA** induces a nonspecific enhancement of learning, we showed that leaky or induced expression of *hsp-PKA** in the wild-type background did not increase the learning score even if flies were undertrained (Fig. 2b). For under-training, flies were subjected to 3 repeats of electric shock in a training trail instead of 12 (see Methods, and ref. 16). We conclude that the PKA* effect is not nonspecific and that the learning defect observed in *NF1* mutants can be rescued by induction of PKA activity. Therefore, the biochemical deficiency in the *NF1* mutants must reside upstream of PKA induction in the cAMP pathway.

These behavioural analyses corroborate previously reported electrophysiological data¹³ that indicated that NF1 might exert its effect through regulation of the activation of Rut-AC. Biochemical assays provide direct evidence to support the idea. Previous experiments have shown that Rut-AC expressed in a cell line can be stimulated not only by Ca²⁺/calmodulin, but also by reagents that stimulate G-proteins, including GTPγS and AIF₄⁻ (see ref. 22). We first examined AC activity in membrane fractions of adult brain tissues. The basal level of AC activity was very similar in the control (K33) and *NF1* mutant membranes, but the GTPγS-stimulated AC activity was markedly reduced in *NF1^{P1}* and *NF1^{P2}* mutant membranes (Fig. 3a). However, significant GTPγS-stimulated activity occurred above the basal level in the mutants. Overexpression of NF1 in control flies did not increase AC activity (data not shown), whereas the reduction in stimulated AC activity seen in *NF1* mutants was mostly rescued by acutely induced expression of the *NF1* transgene (Fig. 3a), indicating that NF1 is indeed able to regulate cAMP synthesis. Thus, GTPγS-stimulated AC activity consists of two components: one that is NF1 dependent and one that is NF1 independent. To determine whether the NF1-dependent AC activity is due to Rutabaga, we assayed *rut¹* and *rut¹;NF1^{P2}* mutant flies. The basal and GTPγS-stimulated level of AC activity were very similar in the single mutant, *rut¹*, and in the double mutant, *rut¹;NF1^{P2}* (Fig. 3b). Thus, the *NF1* mutation has no impact on AC activity in the absence of Rut-AC. In other words, the NF1-dependent cAMP activity is mediated through Rut-AC.

We also measured the Ca²⁺ dependence of the NF1 effect to determine how Rut-AC is involved. Membrane fractions extracted from abdominal tissues were used because the Ca²⁺-dependent Rut-AC activity is easier to detect (Fig. 3c, d). Our data are consistent with a previous report¹⁴ that the Ca²⁺-dependent peak of AC activity is missing in *rut* mutants. Again, the *NF1* mutation had no effects on AC activity across Ca²⁺ concentrations without Rut-AC (Fig. 3d). Moreover, the Ca²⁺-dependent peak of Rut-AC activity was dependent on G-protein stimulation but not on the presence of NF1 (Fig. 3c).

Together, our results reveal a new mechanism for how G-proteins activate the cAMP pathway for normal learning and memory. The G-protein-activated AC activity is both NF1 dependent and NF1 independent. The NF1-dependent component involves Rut-AC. One possibility for how NF1 regulates AC activity is that NF1 acts as a GAP not only for the small G-protein Ras but also for heterotrimeric G-proteins. Thus, NF1 is required for a functional interaction between AC and heterotrimeric G-proteins, similar to the involvement of IRA, a Ras-GAP that is distantly related to NF1, in the Ras-activated cAMP pathway in yeast²⁶.

Another possibility is that NF1 regulates AC activity independently of its role as a Ras-GAP. Therefore, NF1 may be important for coordinating activities of multiple signal-transduction pathways. Nevertheless, the expression of the tumour-suppressor gene *NF1* and its regulation of the Rut-AC signal-transduction pathway are critical to the biochemical processes underlying olfactory learning in *Drosophila*. Similar mechanisms are to be expected in vertebrates. □

Methods

Fly stocks

NF1^{P1}, *NF1^{P2}* and K33 flies have a similar genetic background¹² and were outcrossed with *w¹¹¹⁸* (*isoCJ1*), an isogenic line²¹, for five generations. Thus, *NF1^{P1}*, *NF1^{P2}* and K33^u have a genetic background that is similar to *w¹¹¹⁸* (*isoCJ1*). The transgenic *hsNF1* gene is inserted in the second chromosome¹². In all experiments related to *hsNF1*, only heterozygous *hsNF1/+* was used, which avoided any recessive effects of the insertion on behaviour. *hsp70-PKA** flies have a murine PKA transgene, with His87Gln and Trp196Arg substitutions that prevent interaction with the PKA regulatory subunit²⁵.

Pavlovian learning

Flies were trained by exposure to electroshock (12 pulses at 60 V, duration 1.5 s, interval 5 s) paired with one odour (benzaldehyde (BA, 4%) or methycyclohexanol (MCH, undiluted) for 60 s) and subsequent exposure to the second odour without electroshock. The odour concentrations were adjusted to assume no preference for flies exposed simultaneously to the two odours before the training. Immediately after training, learning was measured by allowing flies to choose between the two odours used during training. No preference between odours results in zero (no learning) PI. Avoidance of the odour previously paired with electroshock produces a 0 < PI ≤ 1.00 (see ref. 16).

Olfactory acuity

Absolute odour avoidance responses were quantified by exposing naive flies to each odour (BA or MCH) or air in the T-maze. After 120 s, the numbers of flies in each arm of the T-maze were counted, and the performance index was calculated for each odour individually as reported²⁰.

Shock reactivity

The ability to sense and escape from electric shock was quantified by inserting electrifiable grids into both arms of the T-maze, and delivering shock pulses to one arm. Flies were transported to the choice point of the T-maze, where they could choose between the two arms. After 60 s, the centre compartment was closed, trapping flies in their respective arms. Individual PI was calculated as defined²⁰.

Adenylyl cyclase activity assay

The described AC activity assay²⁷ was modified as follows. The membrane fraction was extracted from either male heads, with the cuticle removed to leave essentially brain tissue, or from 10 male abdomens. Twenty dissected brains in 850 μl lysis buffer were homogenized, and the membrane fraction was extracted by centrifugation at 178,000g for 10 min. GTPγS (final concentration, 20 μM) was added to stimulate G-protein-activated AC activity immediately after homogenization but before centrifugation. When assaying the basal level of AC activity, GTPγS was not added. These modifications seemed to amplify the NF1 effect. Calcium concentrations were calculated according to MaxChelator v1.31 (see ref. 28).

RT-PCR of induced NF1 expression

Total RNA was prepared from 100 mg of flies using the RNeasy Mini Kit (Qiagen). We isolated mRNA from total RNA (20–100 μg) by using either the Dynabeads mRNA DIRECT Micro Kit or the Dynabeads mRNA Purification Kit (Dyna) and first-strand complementary DNA was synthesized directly from this mRNA using the SuperScript Preamplification System (Gibco BRL). PCR amplification with *Taq* DNA polymerase (Gibco BRL) was carried out in the manufacturer's buffer using 1.5 mM MgCl₂ with 30 cycles of 94°C 1 min, 55°C 1 min and 72°C 1 min, followed by extension at 72°C for 10 min. NF1-specific primers, 5'-tcaccaagctcagagcagca-3' and 5'-gcctgttgccaggttgatg-3', were designed to amplify a 1,251-base-pair (bp) region of cDNA between bases 2,730 and 3,981 (Genbank L26501)¹². These primers span a region of genomic DNA containing a 54-bp intron that allows distinction between cDNA products and potential genomic contaminants. Lack of genomic DNA contamination was confirmed by the absence of any PCR products when RT-PCR was carried out in control reactions without reverse transcriptase. Ribosomal protein rp49-specific primers, 5'-atgaccatccgccagcagc-3' and 5'-gagaacgcaggcagcgttgg-3', were designed to amplify a 391-bp fragment between bases 1 and 391 of the rp49 coding region (Genbank Y13939)²⁹. The control rp49 mRNA should be expressed at equal levels in all cells at all stages²⁹.

Received 28 September; accepted 19 November 1999.

1. Ballester, R. *et al.* The NF1 locus encodes a protein functionally related to mammalian GAP and yeast IRA proteins. *Cell* **63**, 851–859 (1990).
2. Bernards, A. Neurofibromatosis type 1 and Ras-mediated signaling: filling in the GAPs. *Biochim. Biophys. Acta* **1242**, 43–59 (1995).
3. Huson, S. M., Harper, P. S. & Compston, D. A. Von Recklinghausen neurofibromatosis. A clinical and population study in south-east Wales. *Brain* **111**, 1355–1381 (1988).
4. McCormick, F. Ras signaling and NF1. *Curr. Opin. Genet. Dev.* **5**, 51–55 (1995).
5. Xu, G. F. *et al.* The catalytic domain of the neurofibromatosis type 1 gene product stimulates ras GTPase and complements ira mutants of *S. cerevisiae*. *Cell* **63**, 835–841 (1990).
6. Lowy, D. R. & Willumsen, B. M. Function and regulation of Ras. *Annu. Rev. Biochem.* **62**, 851–891 (1993).
7. Jacks, T. *et al.* Tumour predisposition in mice heterozygous for a targeted mutation in NF1. *Nature Genet.* **7**, 353–361 (1994).
8. Ferner, R. E., Hughes, R. A. & Weinman, J. Intellectual impairment in neurofibromatosis 1. *J. Neurol. Sci.* **138**, 125–133 (1996).

9. North, K., Joy, P., Yuille, D., Cocks, N. & Hutchins, P. Cognitive function and academic performance in children with neurofibromatosis type 1. *Dev. Med. Child Neurol.* **37**, 427–436 (1995).
10. North, K. *et al.* Specific learning disability in children with neurofibromatosis type 1: significance of MRI abnormalities. *Neurology* **44**, 878–883 (1994).
11. Silva, A. J. *et al.* A mouse model for the learning and memory deficits associated with neurofibromatosis type 1. *Nature Genet.* **15**, 281–284 (1997).
12. The, I. *et al.* Rescue of a *Drosophila* NF1 mutant phenotype by protein kinase A. *Science* **276**, 791–794 (1997).
13. Guo, H. F., The, I., Hannan, F., Bernards, A. & Zhong, Y. Requirement of *Drosophila* NF1 for activation of adenylyl cyclase by PACAP38-like neuropeptide. *Science* **276**, 795–798 (1997).
14. Livingstone, M. S., Sziber, P. P. & Quinn, W. G. Loss of calcium/calmodulin responsiveness in adenylyl cyclase of *rutabaga*, a *Drosophila* learning mutant. *Cell* **37**, 205–215 (1984).
15. Han, P.-L., Levin, L. R., Reed, R. R. & Davis, R. L. Preferential expression of the *Drosophila* rutabaga gene in mushroom bodies, neural centers for learning in insects. *Neuron* **9**, 619–627 (1992).
16. Tully, T. & Quinn, W. G. Classical conditioning and retention in normal and mutant *Drosophila melanogaster*. *J. Comp. Physiol. A Sens. Neural. Behav. Physiol.* **157**, 263–277 (1985).
17. de Belle, J. S. & Heisenberg, M. Associative odor learning in *Drosophila* abolished by chemical ablation of mushroom bodies. *Science* **263**, 692–695 (1994).
18. Connolly, J. B. *et al.* Associative learning disrupted by impaired Gs signaling in *Drosophila* mushroom bodies. *Science* **274**, 2104–2106 (1996).
19. Grotewiel, M. S., Beck, C. D., Wu, K. H., Zhu, X. R. & Davis, R. L. Integrin-mediated short-term memory in *Drosophila*. *Nature* **391**, 455–460 (1998).
20. Dura, J.-M., Preat, T. & Tully, T. Identification of linotte, a new gene affecting learning and memory in *Drosophila melanogaster*. *J. Neurogenetics* **9**, 1–14 (1993).
21. Yin, J. C. P. *et al.* Induction of a dominant negative CREB transgene specifically blocks long-term memory in *Drosophila*. *Cell* **79**, 49–58 (1994).
22. Levin, L. R. *et al.* The *Drosophila* learning and memory gene rutabaga encodes a Ca^{2+} /calmodulin-responsive adenylyl cyclase. *Cell* **68**, 479–489 (1992).
23. Byers, D., Davis, R. L. & Kiger, J. A. Defect in cyclic-AMP phosphodiesterase due to the dunce mutation of learning in *Drosophila melanogaster*. *Nature* **289**, 79–81 (1981).
24. Chen, C. N., Denome, S. & Davis, R. L. Molecular analysis of cDNA clones and the corresponding genomic coding region of the *Drosophila* dunce⁺ locus, the structure gene for cAMP phosphodiesterase. *Proc. Natl Acad. Sci. USA* **86**, 3599–3603 (1986).
25. Jiang, J. & Struhl, G. Protein kinase A and hedgehog signaling in *Drosophila* limb development. *Cell* **80**, 563–572 (1995).
26. Mitts, M. R., Bradshaw-Rouse, J. & Heideman, W. Interactions between adenylyl cyclase and the yeast GTPase-activating protein IRA1. *Mol. Cell. Biol.* **11**, 4591–4598 (1991).
27. Livingstone, M. S. Genetic dissection of *Drosophila* adenylyl cyclase. *Proc. Natl Acad. Sci. USA* **82**, 5992–5996 (1985).
28. Bers, D. M., Patton, C. W. & Nuccitelli, R. A practical guide to the preparation of Ca^{2+} buffers. *Methods. Cell. Biol.* **40**, 3–29 (1994).
29. O'Connell, P. & Rosbash, M. Sequence, structure and codon preference of the *Drosophila* ribosomal protein 49 gene. *Nucleic Acids Res.* **12**, 5495–5513 (1984).

Acknowledgements

We thank T. Tully for help with behavioural assays, R. Davis for extensive comments on the manuscript, and A. Bernards for providing the fly stocks. This work was supported by a Pew Scholarship, grants from NIH, grants from Texas Neurofibromatosis foundation, NF Inc. Mass Bay Aere and Illinois NF Inc., Perkin fund, and donations from M. L. Rankowitz and S. H. Heffron to Y.Z., and by a National Neurofibromatosis Foundation Young Investigator Award to F.H.

Three distinct and sequential steps in the release of sodium ions by the Na^+/K^+ -ATPase

Miguel Holmgren*, Jonathan Wagg*, Francisco Bezanilla*, Robert F. Rakowski*, Paul De Weer* & David C. Gadsby*

The Marine Biological Laboratory, Woods Hole, Massachusetts 02543, USA

The Na^+/K^+ pump, a P-type ion-motive ATPase, exports three sodium ions and then imports two potassium ions in each transport cycle. Ions on one side of the membrane bind to sites within the protein and become temporarily occluded (trapped within the protein) before being released to the other side^{1,2}, but

details of these occlusion and de-occlusion transitions remain obscure for all P-type ATPases. If it is deprived of potassium ions, the Na^+/K^+ pump is restricted to sodium translocation steps³, at least one involving charge movement through the membrane's electric field^{4,5}. Changes in membrane potential alter the rate of such electrogenic reactions and so shift the distribution of enzyme conformations. Here we use high-speed voltage jumps to initiate this redistribution and show that the resulting pre-steady-state charge movements relax in three identifiable phases, apparently reflecting de-occlusion and release of the three sodium ions. Reciprocal relationships among the sizes of these three charge components show that the three sodium ions are de-occluded and released to the extracellular solution one at a time, in a strict order.

The main electrical signals generated during Na^+/K^+ pumping result from Na^+ traversing part of the membrane's electric field in an access channel that connects Na^+ -binding sites to the extracellular medium^{6–9}. To investigate de-occlusion and release of the three transported Na^+ ions, we measured pump-mediated charge translocation in voltage-clamped, internally dialysed squid giant axons, using solutions designed to limit Na^+/K^+ pumps to phosphorylated conformations with Na^+ -binding sites either occupied and buried or open to the external solution: $(\text{Na}_3)\text{E}_1\text{-P} \leftrightarrow \text{P-E}_2\text{-Na}_3 \leftrightarrow \text{P-E}_2$ (Fig. 1, dotted box). Pump-mediated charge was assayed as the component of membrane current that was sensitive to dihydrodigeitoxigenin (H_2DTG), a specific Na^+/K^+ -pump inhibitor¹⁰ (Fig. 2a (2–3)). With 100 mM external Na^+ ($[\text{Na}]_o$), the change in pump current induced by a voltage jump (20-ms step from 0 mV to –90 mV; current displayed at 50 μs per point in Fig. 2a) comprised fast (the first ~10 points) and slow ($\tau \approx 4$ ms) components, and relaxed to near zero. As negative internal potentials electrostatically favour the approach of external Na^+ to their binding sites within the pump, we interpret the fast component (see ref. 8) as reflecting rapid electrogenic binding of Na_o^+ to vacant sites (P-E_2) to satisfy the new $\text{P-E}_2 \leftrightarrow \text{P-E}_2\text{-Na}_3$ distribution demanded by the new membrane potential, –90 mV. The resulting increased abundance of pumps in the $\text{P-E}_2\text{-Na}_3$ state drives the slower (comparatively electroneutral; see below) occlusion reaction ($\text{P-E}_2\text{-Na}_3 \leftrightarrow (\text{Na}_3)\text{E}_1\text{-P}$), which is tracked by further electrogenic binding of Na_o^+ to P-E_2 , now rate limited by the slow conformational change and hence appearing as slow charge movement (see, for example, refs 7, 11). We found no recruitment of additional slow charge (Q_s ; Fig. 2b, left) at extreme potentials, in accord with the expectation that, in the steady state, pumps in the P-E_2 conformation should be

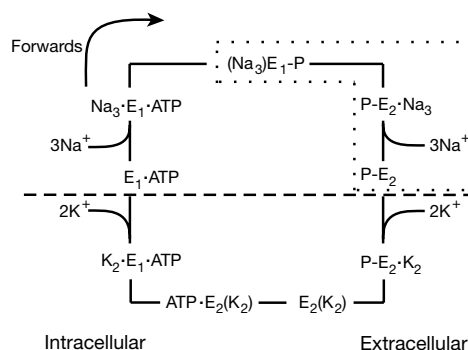


Figure 1 Simplified Post-Albers transport cycle emphasizing two principal Na^+/K^+ pump conformations: E_1 with ion-binding sites facing the cytoplasm, and E_2 with ion-binding sites open to the extracellular solution. Phosphorylation of E_1 by ATP occludes three Na^+ , which are released to the external medium after the conformational change to E_2 , whereupon two K^+ bind, eliciting dephosphorylation and K^+ occlusion. ATP binding favours transition back to E_1 , prompting K^+ release to the cytoplasm and binding of three Na^+ , completing the cycle. The dashed (horizontal) line separates Na^+ - and K^+ -translocation pathways; the dotted box encloses states isolated experimentally to yield the charge movements examined.

* Present addresses: Department of Neurobiology, Harvard Medical School, Boston, Massachusetts 02115, USA (M.H.); rSafe Inc, 369 Pine St, San Francisco, California 94104, USA (J.W.); Department of Physiology, UCLA School of Medicine, Los Angeles, California 90095, USA (F.B.); Department of Physiology and Biophysics, FUHS/Chicago Medical School, North Chicago, Illinois 60064, USA (R.F.R.); Department of Physiology, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania 19104, USA (P.D.); Laboratory of Cardiac Membrane Physiology, The Rockefeller University, New York, New York 10021, USA (D.C.G.).

scarce at very negative potentials but maximally abundant at very positive potentials. Also as expected, the positive and negative potentials at which slow charge recruitment effectively ceased depended on $[Na]_o$ (Fig. 2b, left): at lower $[Na]_o$ less positive potentials can fully unload Na^+ -binding sites, and at higher $[Na]_o$ less negative potentials are required for Na^+ to fill all P-E₂ sites.

The relaxation rate (k) of the slow component reached a minimum ($\sim 100\text{ s}^{-1}$) at positive potentials, but became faster at negative potentials, and tended towards saturation (at $\sim 1,400\text{ s}^{-1}$) at large

negative voltages (Fig. 2b, right). This may indicate that, ultimately, regardless of $[Na]_o$, the reactions limiting the rates of the slow charge translocation in both the forwards (de-occlusion: extreme positive potentials) and backwards (occlusion: extreme negative potentials) directions are nearly electroneutral^{7,8,12}. Between these extremes the slow relaxation rate varies with $[Na]_o$ and voltage, reflecting the filling of Na^+ -binding sites¹¹. Fitting these slow rates (curves, Fig. 2b, right) to the simple three-state scheme (Fig. 1, box) gave an estimate for the fraction (λ) of the membrane's electric field dropped along the access channel of 71% similar to our previous estimate (69%) from analysis of the voltage dependence of $^{22}Na^+$ efflux mediated by Na_i/Na_o exchange under similar conditions⁷.

We tested whether the reactions underlying the fast and slow components of pump charge movement occur in sequence, as implied above, by making the membrane potential negative enough (-110 mV) for long enough (20 ms) for complete relaxation of the current in 400 mM $[Na]_o$ before stepping back to 0 mV. As anticipated, this conditioning step elicited a large fast component (reflecting Na^+ binding) followed by a slow relaxation (reflecting occlusion); but, although the jump back to 0 mV caused the return movement of a comparable amount of slow charge, the fast component preceding it was greatly reduced (Fig. 3a). Evidently, because the occlusion/de-occlusion reaction is strongly poised towards occlusion at -110 mV in high $[Na]_o$ (Fig. 2b, left), the prolonged sojourn at this potential drove most pumps to the occluded state, $(Na_3)E_1-P$, whence they could not immediately release Na^+ (to generate rapid charge movement) when the potential was returned to 0 mV. Positive voltage jumps also elicit fast and slow movements of pump-mediated charge (Fig. 3b), the fast component in this case reflecting migration of released Na^+ ions out through the access channel, and the slow component tracking the rate-limiting de-occlusion step. The prolonged positive potential should thus leave newly unloaded pumps available to bind Na^+ and, hence, to contribute to a slightly larger fast charge component upon the jump back to 0 mV (as observed, Fig. 3b). These findings with long conditioning voltage steps show that the reactions underlying the fast and slow charge movements are obliged to occur in sequence.

The simple scheme of Fig. 1 cannot, however, explain the comparable amounts of slow charge moved during and after long conditioning voltage jumps (for example, Figs 2a, 3a), particularly large negative jumps⁸. If the three Na^+ bind virtually instantaneously to P-E₂, as implied (Fig. 1), a sufficiently negative voltage jump should rapidly saturate available Na^+ -binding sites, causing a maximally large movement of fast-charge. But, if the

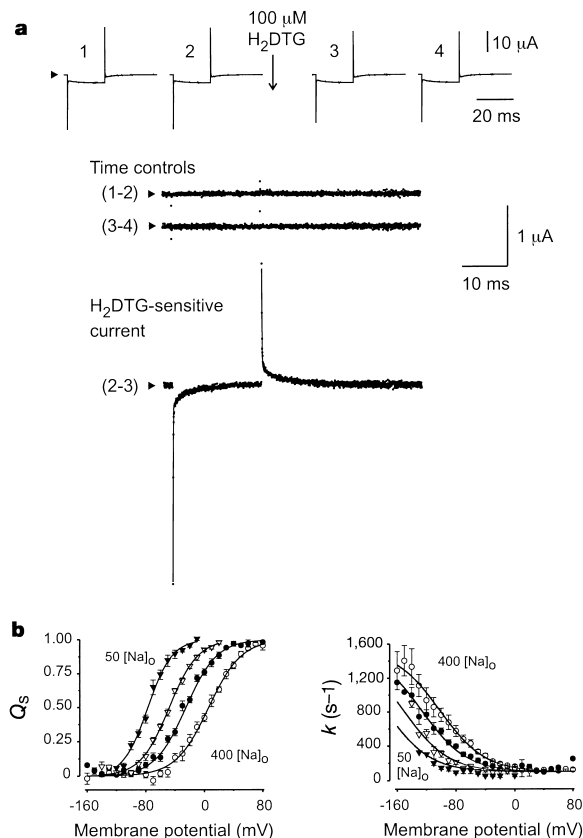


Figure 2 Charge movement during Na^+ translocation by the Na^+/K^+ pump. **a**, Determination of H₂DTG-sensitive (pump-mediated) current. Traces 1–4 show averages of 100 current responses to 20-ms voltage jumps, from 0 to -90 mV , applied at 4-min intervals to an axon exposed to 100 mM $[Na]_o$ without (1 and 2) or with (3 and 4) 100 μM H₂DTG. Difference currents (1–2) and (3–4) show stability with time; (2–3) shows pump-mediated current fitted with a two-exponential function (solid lines largely obscured by data points), yielding τ values of 0.06 ms and 4 ms at -90 mV (on), and 0.1 ms and 5 ms at 0 mV (off). Arrowheads mark zero current. Filter 5 kHz, sampling 20 kHz. **b**, Voltage dependence of slow component, elicited by 20–40-ms jumps from -40 mV at 400 (open circles), 200 (filled circles), 100 (open triangles) or 50 (filled triangles) mM $[Na]_o$. Left, normalized slow charge (Q_s); Boltzmann fits yield, at 400, 200, 100 or 50 mM $[Na]_o$, (equivalent valence) $z_q = 1.15 \pm 0.05$, 1.22 ± 0.04 , 1.32 ± 0.04 or 1.55 ± 0.09 , and (midpoint potential) $V_q = 6 \pm 1$, -25 ± 1 , -48 ± 1 or $-77 \pm 1\text{ mV}$. Right, relaxation rate constant (k); curves show simultaneous fit to modified Hill expression⁷:

$$k_{(V)} = k_f + \frac{k_b}{1 + \frac{\{K_{0.5}(0) \exp(\lambda VF/RT)\}^n}{[Na]_o^n}}$$

where k_f and k_b are voltage-independent forwards (de-occlusion) and backwards (occlusion) rate constants, $K_{0.5}(0)$ is half-activating $[Na]_o$ at 0 mV, n is the Hill coefficient and λ is the fraction of the electrical field dropped along the access channel. Overall best-fit values: $k_f = 97 \pm 12\text{ s}^{-1}$, $k_b = 1,444 \pm 367\text{ s}^{-1}$, $K_{0.5}(0) = 6.95 \pm 1.47\text{ M}$, $n = 1.17 \pm 0.1$ and $\lambda = 0.71 \pm 0.04$.

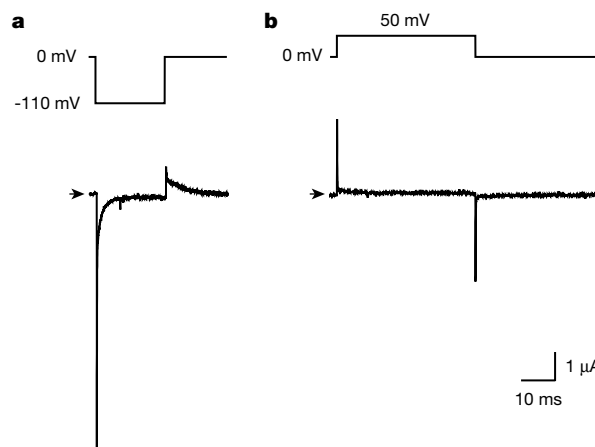


Figure 3 Non-conservation of fast charge between on and off voltage steps. **a**, H₂DTG-sensitive current at 400 mM $[Na]_o$ due to test step to -110 mV , with on-line subtraction of scaled linear capacitance current elicited by small steps from the steady potential, -110 mV . Filter 20 kHz, sampling 100 kHz. **b**, H₂DTG-sensitive current in the same axon with same conditions as **a**, but for step to $+50\text{ mV}$. Filter 20 kHz, sampling 50 kHz.

$P-E_2\cdot Na_3 \leftrightarrow (Na_3)E_1-P$ occlusion reaction is intrinsically nearly electroneutral, as concluded, the subsequent slow conformational transition should engender only a negligible electrical signal, in contrast to the large movement of slow charge recorded (Fig. 3a). One proposal to account for such a discrepancy makes the access channel state a high-energy, short-lived conformation sandwiched between two occlusion/de-occlusion transitions⁸. This effectively limits the size of the rapid charge movement (because the access-channel state is rarely populated), but preserves a large slow component (because release and binding of external Na^+ must still proceed through the access channel). Our extension of the model (Fig. 4a) posits that the three Na^+ within $(Na_3)E_1-P$ are de-occluded and released one at a time (see also ref. 12). Two predictions of this new scheme are that the relaxation of pump current after a voltage jump should comprise more than two components, and that reciprocal relationships should exist between the magnitudes of those components.

To resolve additional components we increased the sampling rate to 2 MHz and used both briefer voltage steps and low $[Na]_o$ to avoid driving most pumps to the occluded state. Figure 4b shows the pump current elicited by a 500- μ s jump from 0 to -110 mV at 100 mM $[Na]_o$. Under these conditions the slow relaxation ($\tau \approx 3$ ms; Fig. 2b), appearing as an almost steady current at -100 mV, can move only $\sim 15\%$ of its charge during the 500- μ s pulse. Because it moves even more slowly at 0 mV ($\tau \approx 5$ ms; Fig. 2b), restoration of that small slow charge after the jump back to 0 mV is barely discernible as a small (practically time-independent) displacement above the zero-current line (Fig. 4b).

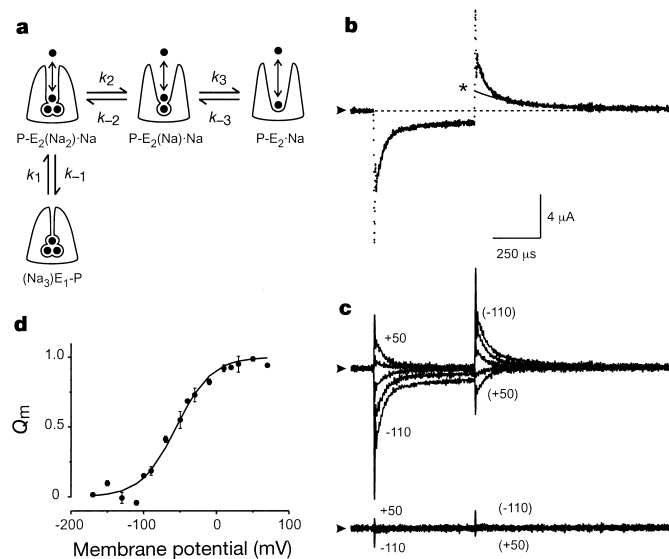


Figure 4 Faster sampling reveals medium-speed component. **a**, Model: predominant electrogenic events are assumed to be migration of released Na^+ (black balls) along access channels, narrow (high-field) for the first Na^+ but wide (low-field) for the next two Na^+ ; relatively electroneutral transitions (rate constants k_1/k_{-1} , k_2/k_{-2} and k_3/k_{-3}) de-occlude each Na^+ . **b**, Pump current elicited by 500- μ s step from 0 to -110 mV, at 100 mM $[Na]_o$; solid line (asterisk) shows 2-exponential fit (ignoring first 150 μ s) to off current; the medium-speed component, $\tau = 220$ μ s, was >20 -fold faster than the slow component, but ~ 10 -fold slower than the fast component(s) which relaxed in two phases (above fit line) with $\tau \approx 3$ μ s and ≈ 30 μ s (closely matching two phases of capacitance current relaxation in that axon, and so reflecting the speed of membrane potential change). Filter 100 kHz, sampling 2 MHz. **c**, Superimposed pump currents for steps from 0 mV to +50, +10, -30, -70 and -110 mV at 100 mM $[Na]_o$; below, time controls show subtraction of records taken 4 min apart before adding H_2DTG . Filter 100 kHz, sampling 2 MHz. **d**, Voltage sensitivity of normalized medium-speed charge (Q_m ; mean, $n = 3$) at 50 mM $[Na]_o$, measured at 0 mV (as in **b**) after 500- μ s steps to indicated potentials; Boltzmann fit parameters, $z_q = 1.1 \pm 0.1$ and $V_q = -55 \pm 2$ mV.

The large-amplitude charge movements (two phases, mirroring the two phases of capacitance current decline; see Fig. 4b) that decay within the first ~ 100 μ s after both negative- and positive-going voltage steps correspond to the fast component already described (Figs 2a, 3). Then, however, a clear inflection in the current relaxation at 0 mV, ~ 100 μ s after its onset, reveals a subsequent additional, medium-speed component with a time constant of roughly 200 μ s (fitted curve; asterisk, Fig. 4b). The charge moved at 0 mV by this medium-speed component was small following jumps to positive potentials, but, like the slow component, it became prominent after pulses to potentials more negative than -70 mV (Fig. 4c), and its magnitude varied steeply with conditioning voltage (Fig. 4d).

If the underlying reactions are sequential (as proposed, Fig. 4a), each component (fast, medium-speed and slow) must develop at the expense of the preceding one. Figure 5a shows that the growth of the medium-speed component (assayed at 0 mV; fitted curves) with prolongation of the preceding -100-mV conditioning pulse from 200 to 800 μ s was accompanied by a decrease in the amplitude of the fast component. This growth of medium-speed charge (Q_m) with conditioning pulse duration was more rapid at 400 mM than at 100 mM $[Na]_o$ (Fig. 5b), occurring in both cases with a time course approximating the relaxation of the medium-speed component at -100 mV at the same $[Na]_o$. However, further prolongation of the conditioning pulse reduced the magnitude of the subsequent Q_m , with time constants similar to those measured directly for relaxation of the slow charge component at -100 mV and at the appropriate $[Na]_o$ (Fig. 2b). Together, these results support our model (Fig. 4a) by establishing that the medium-speed component grows at the expense of the fast component, but then wanes as it populates the enzyme conformations that generate the slow component.

We have shown here that the charge movements accompanying release of three Na^+ ions by the pump are rate limited by three

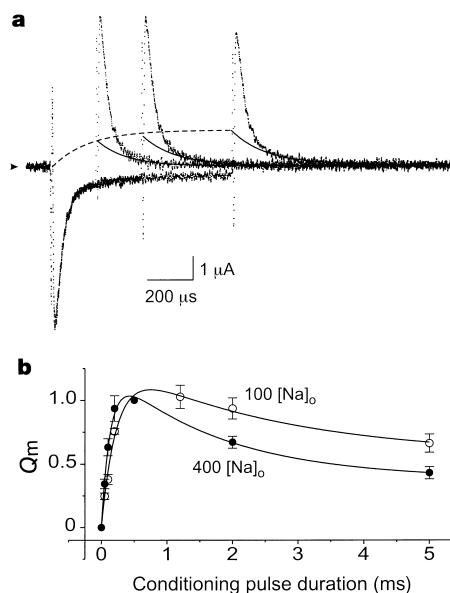


Figure 5 Sequential reactions underlie the three charge components. **a**, Conditioning pulse duration (200, 400 and 800- μ s steps from 0 to -100 mV at 50 mM $[Na]_o$) influences relative sizes of subsequent fast and medium-speed components: solid lines represent 2-exponential fits extrapolated to the off voltage jump; for all three traces, medium-speed $\tau = 140$ μ s. For illustration, an exponential curve (τ also 140 μ s) has been drawn through the extrapolated initial amplitudes. Filter 100 kHz, sampling 2 MHz. **b**, Normalized medium-speed charge (Q_m) moved at 0 mV after steps to -100 mV of indicated duration, at 100 mM (open circles) and 400 mM (filled circles) $[Na]_o$; curves are fits to difference of 2 exponentials, yielding τ values of 0.13 ± 0.01 and 1.74 ± 0.45 ms at 400 mM $[Na]_o$, whereas at 100 mM $[Na]_o$ τ values are 0.23 ± 0.01 and 2.6 ms (the latter constrained as the mean τ for the slow component at -100 mV and 100 mM $[Na]_o$, from Fig. 2b, right).

separate, sequential reactions, each of which probably opens a binding site to permit one Na^+ to escape. Each charge component has well defined characteristics. The slowest appears to reflect strongly electrogenic (equivalent valence, $z \approx 1$; Fig. 2b) release of the first Na^+ , through $\sim 70\%$ of the membrane field, in a reaction that is rate limited by the slow (~ 100 to $1,400 \text{ s}^{-1}$) major $\text{E}_1\text{-P} \leftrightarrow \text{P-E}_2$ conformational change, which itself seems relatively electroneutral; this slow component shows low sensitivity to $[\text{Na}]_o$ (Fig. 2b) but has strongly temperature-sensitive rates⁵, revealing an enthalpic activation energy of $\sim 80 \text{ kJ mol}^{-1}$ ($10\text{--}20^\circ\text{C}$; not shown; see ref. 13). Like the slow component, the medium-speed ($\sim 6,000$ to $20,000 \text{ s}^{-1}$) component also has a steeply voltage-dependent charge magnitude ($z \approx 1$; Fig. 4d), and a relaxation rate that increases with $[\text{Na}]_o$ at negative potentials and shows a high activation energy ($\sim 70 \text{ kJ mol}^{-1}$; not shown) and, hence, probably reflects the reaction that de-occludes the second Na^+ . The fast component mirrors the time course of the distributed membrane capacitance transient and so must reflect charge transitions with rates $\geq 10^6 \text{ s}^{-1}$, appropriate for rapid Na^+ release through an access channel, but still possibly rate limited by a minor conformational change that de-occludes the final Na^+ ; consistent with their high speed, these relaxations show little temperature sensitivity (not shown). In further contrast to the slow and medium-speed components, the fast charge movement has extremely weak voltage sensitivity (note simple scaling of the fast component amplitude with potential over a 160-mV range in Fig. 4c), and it is seen in virtual isolation at very low $[\text{Na}]_o$ ($\leq 25 \text{ mM}$; not shown), indicating that it may reflect release of the final Na^+ ion(s) from a relatively high-affinity site(s) on P-E_2 (see refs 8, 9, 12). Our failure to observe any comparably high-speed charge movement displaying the strong voltage sensitivity of the medium-speed and slow components, despite exploring a broad range of $[\text{Na}]_o$ and voltage, argues (see ref. 8) that there must be negligible steady-state occupancy of the narrow (high-field) access-channel conformation $\text{P-E}_2(\text{Na}_2)\text{-Na}$, which we propose (Fig. 4a) is ultimately responsible for those slower charge relaxations; this in turn implies that both rate constants leading away from that state (k_{-1} and k_2 in Fig. 4a) are relatively large.

The strictly sequential nature of the three charge components shown here indicates that the three Na^+ may be released from the Na^+/K^+ pump in a fixed order. Ordered occlusion/de-occlusion of two K^+ by kidney microsomal Na^+/K^+ -ATPase¹⁴ and sequential occlusion, translocation and release of the two Ca^{2+} ions transported by the sarcoplasmic reticulum Ca^{2+} -ATPase¹⁵ have been detected using isotopes and rapid filtration techniques (time resolution $\sim 10 \text{ ms}$), but the far higher time resolution and sensitivity of the electrical recording methods used here permit extraction of finer molecular kinetic detail^{8,12,16}. Closer examination, using these methods, of the interactions of extracellular Na^+ ions with their binding sites within the Na^+/K^+ pump will now be required to discern the precise molecular rearrangements that surround these principal charge movements in the Na^+/K^+ transport cycle. □

Methods

Giant axons from the squid *Loligo pealei* were voltage clamped¹⁷, internally dialysed and externally superfused at $20\text{--}22^\circ\text{C}$ with Cl^- -free solutions^{7,10} designed to restrict the pump to Na^+ de-occlusion/release steps (Fig. 1). Intracellular (in mM; pH adjusted with HEPES): 80 Na-HEPES, 57 N-methyl-D-glucamine(NMG)-HEPES, 50 glycine, 50 phenylpropyltriethylammonium-sulphate, 5 dithiothreitol, 2.5 1,2-bis(2-aminophenoxy)ethane-N,N,N',N'-tetraacetic acid (BAPTA), 15 Mg-HEPES, 5 Tris-ATP, 5 phospho(enol)pyruvate tri- Na^+ -salt and 5 phospho-L-arginine mono- Na^+ -salt. Extracellular (in mM): 400 Na-isethionate, 75 Ca-sulphamate, 1 3,4-diaminopyridine, 2×10^{-4} tetrodotoxin, 5 Tris-HEPES and 0.05 EDTA (pH 7.7). Osmolality of all solutions was $\sim 930 \text{ mOsmol kg}^{-1}$. To lower $[\text{Na}]_o$, Na-isethionate was replaced by tetramethylammonium-sulphamate or NMG-sulphamate. Voltage pulses were generated and currents recorded using a 16-bit PC44 A-D/D-A converter board (Innovative Technologies) with software developed in-house. Currents were filtered at 12.5–200 kHz, then sampled at 20 kHz–2 MHz. Current records were sometimes acquired after subtraction of appropriately amplified small current signals, obtained in a voltage range where pump-mediated

charge movement tended towards saturation, to minimize currents from linear membrane capacitance. Pump current was determined as current sensitive to $100 \mu\text{M H}_2\text{DTG}^{10}$.

Received 22 September; accepted 17 December 1999.

- Post, R. L., Hegyvary, C. & Kume, S. Activation by adenosine triphosphate in the phosphorylation kinetics of sodium and potassium ion transport adenosine triphosphatase. *J. Biol. Chem.* **247**, 6350–6540 (1972).
- Beaugé, L. A. & Glynn, I. M. Occlusion of K ions in the unphosphorylated sodium pump. *Nature* **280**, 510–512 (1979).
- Garrahan, P. J. & Glynn, I. M. The behaviour of the sodium pump in red cells in the absence of external potassium. *J. Physiol.* **192**, 161–174 (1967).
- Fendler, K., Grell, E., Haubs, M. & Bamberg, E. Pump currents generated by the purified Na^+/K^+ -ATPase from kidney on black lipid membranes. *EMBO J.* **4**, 3079–3085 (1985).
- Nakao, M. & Gadsby, D. C. Voltage dependence of Na translocation by the Na/K pump. *Nature* **323**, 628–630 (1986).
- Läuger, P. *Electrogenic Ion Pumps* (Sinauer, Sunderland, MA, 1991).
- Gadsby, D. C., Rakowski, R. F. & De Weer, P. Extracellular access to the Na,K pump: pathway similar to ion channel. *Science* **260**, 100–103 (1993).
- Hilgemann, D. W. Channel-like function of the Na,K pump probed at microsecond resolution in giant membrane patches. *Science* **263**, 1429–1432 (1994).
- Heyse, S., Wuddel, I., Apell, H. J. & Stürmer, W. Partial reactions of the Na,K-ATPase: determination of rate constants. *J. Gen. Physiol.* **104**, 197–240 (1994).
- Rakowski, R. F., Gadsby, D. C. & De Weer, P. Stoichiometry and voltage-dependence of the sodium pump in voltage-clamped, internally-dialyzed squid giant axon. *J. Gen. Physiol.* **93**, 903–941 (1989).
- Läuger, P. & Apell, H. J. Transient behaviour of the Na^+/K^+ -pump: microscopic analysis of nonstationary ion-translocation. *Biochim. Biophys. Acta* **944**, 451–464 (1988).
- Wuddel, I. & Apell, H. J. Electrogenicity of the sodium transport pathway in the Na,K-ATPase probed by charge-pulse experiments. *Biophys. J.* **69**, 909–921 (1995).
- Friedrich, T. & Nagel, G. Comparison of Na^+/K^+ -ATPase pump currents activated by ATP concentration or voltage jumps. *Biophys. J.* **73**, 186–194 (1997).
- Forbush, B. Rapid release of 42K or 86Rb from two distinct transport sites on the Na,K-pump in the presence of Pi or vanadate. *J. Biol. Chem.* **262**, 11116–11127 (1987).
- Inesi, G. Characterization of partial reactions in the catalytic and transport cycle of sarcoplasmic reticulum ATPase. *J. Biol. Chem.* **262**, 16338–16342 (1987).
- Lu, C.-C. et al. Membrane transport mechanisms probed by capacitance measurements with megahertz voltage clamp. *Proc. Natl Acad. Sci. USA* **92**, 11220–11224 (1995).
- Bezani, F., White, M. M. & Taylor, R. E. Gating currents associated with potassium channel activation. *Nature* **296**, 657–659 (1982).

Acknowledgements

This work was supported by grants from the NIH. M.H. was supported in part by the Grass Foundation, and J.W. held a HHMI Postdoctoral Research Fellowship for Physicians.

Correspondence and requests for materials should be addressed to D.C.G.

(e-mail: gadsby@rockvax.rockefeller.edu), or to M.H.

(e-mail: miguel_holmgren@hms.harvard.edu).

The 21-nucleotide *let-7* RNA regulates developmental timing in *Caenorhabditis elegans*

Brenda J. Reinhart[†], Frank J. Slack[†], Michael Basson[‡], Amy E. Pasquinelli^{*}, Jill C. Bettinger[‡], Ann E. Rougvie[#], H. Robert Horvitz[§] & Gary Ruvkun^{*}

^{*} Department of Molecular Biology, Massachusetts General Hospital, and Department of Genetics, Harvard Medical School, Boston, Massachusetts 02114, USA

[§] Howard Hughes Medical Institute, Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

[#] Department of Genetics, Cell Biology and Development, University of Minnesota, St Paul, Minnesota 55108, USA

[†] These authors contributed equally to this work

The *C. elegans* heterochronic gene pathway consists of a cascade of regulatory genes that are temporally controlled to specify the timing of developmental events¹. Mutations in heterochronic genes cause temporal transformations in cell fates in which stage-specific events are omitted or reiterated². Here we show

[‡] Present addresses: Axys Pharmaceuticals, South San Francisco, California 94080, USA (M.B.); Ernest Gallo Clinic and Research Center, UCSF, Emeryville, California 94608, USA (J.C.B.); Department of MDCB, Yale University, New Haven CT 06520, USA (F.J.S.).

that *let-7* is a heterochronic switch gene. Loss of *let-7* gene activity causes reiteration of larval cell fates during the adult stage, whereas increased *let-7* gene dosage causes precocious expression of adult fates during larval stages. *let-7* encodes a temporally regulated 21-nucleotide RNA that is complementary to elements in the 3' untranslated regions of the heterochronic genes *lin-14*, *lin-28*, *lin-41*, *lin-42* and *daf-12*, indicating that expression of these genes may be directly controlled by *let-7*. A reporter gene bearing the *lin-41* 3' untranslated region is temporally regulated in a *let-7*-dependent manner. A second regulatory RNA, *lin-4*, negatively regulates *lin-14* and *lin-28* through RNA–RNA interactions with their 3' untranslated regions^{3,4}. We propose that the sequential stage-specific expression of the *lin-4* and *let-7* regulatory RNAs triggers transitions in the complement of heterochronic regulatory proteins to coordinate developmental timing.

To identify new heterochronic genes, we carried out a genetic screen for mutations that suppress the synthetic sterile phenotype of a strain bearing the *lin-14*(n179) and *egl-35*(n694) mutations. We separated candidate suppressor mutations from the *lin-14* and *egl-35* mutations and examined each mutant for heterochronic defects. Out of 36 suppressor mutations isolated from animals carrying 44,000 mutagenized haploid genomes, the mutation *n2853* caused the strongest retarded heterochronic defects in a *lin14*(+) background (Fig. 1c; Table 1) and a temperature-sensitive adult lethal phenotype associated with vulval bursting. Another suppressor mutation, *mg279*, failed to complement *n2853* and caused a weak retarded phenotype (Table 1). We genetically mapped *n2853* and *mg279* and found that of the lethal mutations in the same region, *let-7*(*mn112*) (ref. 5) displayed heterochronic (Table 1) and lethal phenotypes (93% lethal, *n* = 60) nearly identical to that of *n2853*. *let-7*(*mn112*) is not temperature sensitive and failed to complement both *n2853* and *mg279*.

The first evidence of a *let-7* heterochronic defect is at the L4-to-adult moult. Hypodermal blast cells normally divide at each larval stage and at the adult stage exit the cell cycle, fuse with neighbouring hypodermal seam cells and generate cuticular alae⁶ (Fig. 1a). In *let-7*(*n2853*) animals, the blast cell lineages were normal through the L3-to-L4 moult, but at the L4-to-adult moult, they reiterated larval patterns of cell division and failed to generate alae (Fig. 1a; Table 1). *let-7*(*n2853*) mutant animals reared at the permissive temperature underwent a supernumerary moult to a fifth larval stage, L5 (56%, *n* = 26). At the L5-to-adult moult, seam cells exited the cell cycle, fused with neighbouring seam cells, and produced alae (100%, *n* = 10 animals). The opposite phenotype resulted from over-

expressing *let-7*. Increasing *let-7* gene dosage on a transgenic array caused hypodermal cells to precociously exit the cell cycle and terminally differentiate after the L3-to-L4 moult (83%, *n* = 18 animals). The opposite heterochronic phenotypes caused by reducing or increasing *let-7* activity indicate that *let-7* may function as a temporal switch between larval and adult fates.

let-7 acts upstream of the heterochronic gene *lin-29*, a zinc-finger transcription factor that specifies adult-specific patterns of cell lineage and cell differentiation^{2,7}. In wild-type animals, LIN-29 protein is expressed during the L4 and adult stages in hypodermal cells⁸. Consistent with the delay of adult differentiation by one stage in *let-7*(*n2853*) animals, LIN-29 expression in the hypodermis of L4 stage *let-7* animals was reduced relative to wild type, but expressed at normal levels at the L5 stage (Fig. 1b–d). Thus, *let-7* is necessary for the upregulation of LIN-29 expression in the hypodermis during the L4 stage, which in turn specifies adult cell fates.

The retarded alae phenotype caused by *let-7* mutations was partially suppressed by precocious mutations in the genes *lin-41*, *lin-42*, *lin-14* and *lin-28* (Table 1). For these epistasis experiments, we used the strong *let-7* allele *mn112*, which by molecular criteria completely eliminates gene function (see below). Mutations in *lin-41* and *lin-42* cause precocious expression of adult fates during late larval stages but do not affect L1 and L2 stage fates^{9,10}. Thus, like *let-7* mutations, *lin-41* and *lin-42* mutations specifically affect late larval stage development and these three genes may function at about the same time during development. The *let-7* retarded heterochronic (Table 1) and lethal (F. J. Slack *et al.*, manuscript in preparation) phenotypes were partially suppressed by *lin-41* and *lin-42* mutations; conversely, the precocious alae phenotypes of *lin-41* and *lin-42* mutants were partially suppressed by a *let-7* mutation (Table 1). Although other interpretations are possible, these data are consistent with a model in which *lin-41* and *lin-42* are negatively regulated by *let-7*. Molecular analysis (see below) suggests that regulation by *let-7* may be direct.

Unlike *let-7* mutations, *lin-14* and *lin-28* mutations affect early larval development², suggesting that *let-7* functions later than *lin-14* and *lin-28*. For example, *lin-28*-null mutants delete L2 fates, but double mutant combinations with *let-7* did not suppress this early defect (Table 1). However, the reiteration of larval fates caused by the *let-7*-null mutation was partially suppressed by the precocious expression of adult fates caused by *lin-28* or *lin-14* null mutations

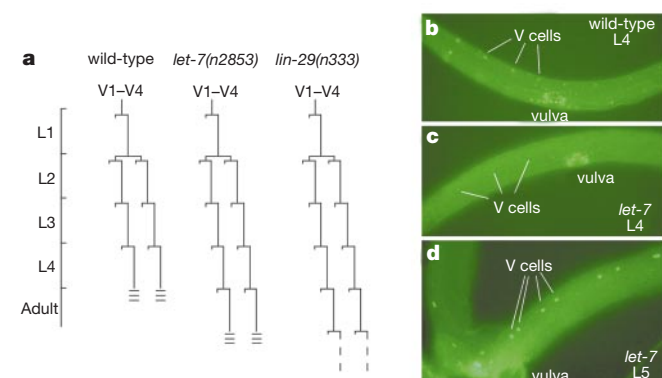


Figure 1 The *let-7* heterochronic phenotype. **a**, Lineage of the lateral hypodermal cells V1, V2, V3 and V4 in wild-type⁶, *let-7*(*n2853*) and *lin-29*(*n333*) animals². L3 and L4 stage V cell lineages are equivalent in hermaphrodites and cannot be distinguished. **b**, Wild-type L4 stage animal with LIN-29 expression in the lateral hypodermis and vulva. **c**, *let-7*(*n2853*) L4 stage animal with LIN-29 expression reduced in V cells but at normal levels in vulval cells. **d**, An L5 stage *let-7*(*n2853*) animal showing accumulation of LIN-29 at high levels one stage later than wild-type animals.

Table 1 Phenotype of *let-7* mutants and interactions with other heterochronic mutants

Strain	Percentage of animals with adult lateral alae*	
	L3 moult	L4 moult
Wild-type (N2)	0 (25)	100 (25)
<i>let-7</i> (<i>mn112</i>) <i>unc-3</i> (<i>e151</i>)	0 (20)	0 (20)
<i>let-7</i> (<i>n2853</i>) 15 °C	0 (20)	0 (20)
<i>let-7</i> (<i>n2853</i>) 25 °C	0 (30)	0 (20)
<i>let-7</i> (<i>mg279</i>)	0 (20)	100† (24)
<i>lin-41</i> (<i>ma104</i>)	54† (48)	100 (45)
<i>lin-41</i> (<i>ma104</i>); <i>let-7</i> (<i>mn112</i>) <i>unc-3</i> (<i>e151</i>)	0 (18)	70 (20)
<i>lin-42</i> (<i>n1089</i>)	90† (72)	100 (58)
<i>lin-42</i> (<i>n1089</i>); <i>let-7</i> (<i>mn112</i>) <i>unc-3</i> (<i>e151</i>)	13 (66)	31 (86)
<i>lin-14</i> (<i>n536</i> <i>n540</i>)	100 (25)	100 (25)
<i>lin-14</i> (<i>n536</i> <i>n540</i>) <i>let-7</i> (<i>mn112</i>) <i>unc-3</i> (<i>e151</i>)	44† (36)	70 (46)
<i>lin-14</i> (<i>n179</i>) 25 °C	100 (38)	100 (28)
<i>lin-14</i> (<i>n179</i>) <i>let-7</i> (<i>mn112</i>) <i>unc-3</i> (<i>e151</i>) 25 °C	18 (32)	30 (129)
<i>lin-28</i> (<i>n719</i>)	100 (20)	100 (20)
<i>lin-28</i> (<i>n719</i>); <i>let-7</i> (<i>mn112</i>); <i>unc-3</i> (<i>e151</i>)†	7† (60)	17† (58)
<i>lin-4</i> (<i>e912</i>)	0 (20)	0 (20)
<i>lin-4</i> (<i>e912</i>); <i>let-7</i> (<i>mn112</i>) <i>unc-3</i> (<i>e151</i>)	0 (20)	0 (40)

All strains were grown at 20° unless otherwise indicated.

* The number of animals is given in parentheses.

† Some animals had patches of alae rather than continuous alae, indicating a mix of larval fates for some V cells and adult fates for others. Percentage of animals with patches: 29% of *let-7*(*mg279*) adults; 25% of *lin-41*(*ma104*) L4s; 28% of *lin-42*(*n1089*) L4s; 33% of *lin-14*(*n536* *n540*) *let-7*(*mn112*) *unc-3*(*e151*) L4s; 7% of *lin-28*(*n719*); *let-7*(*mn112*) *unc-3*(*e151*) L4s; and 17% of *lin-28*(*n719*); *let-7*(*mn112*) *unc-3*(*e151*) adults.

‡ In *lin-28*; *let-7* animals, the postdeirid cells derived from the V5 blast cell during the L2 stage were absent in L4-stage animals (*n* = 5), suggesting that the deletion of V cell L2 fates caused by the *lin-28*(*n719*) null mutation was not suppressed.

(Table 1). This suggests that the early developmental effects of *lin-14* and *lin-28* affect *let-7* function at late larval stages. Because the *let-7* null mutant phenotype is not epistatic to *lin-14* and *lin-28* null mutations, *let-7* is not the only output of these earlier acting genes. Molecular analysis of *let-7* (see below) indicates that direct regulation of *lin-14* and *lin-28* by *let-7* is also possible.

By a combination of transgene complementation, RNA expression analysis and mutant allele sequencing, we established that *let-7* encodes an untranslated RNA. This analysis mapped *let-7* to a 2.5-kilobase (kb) region of cosmid C05G5 that fully complemented *let-7(mn112)* (Fig. 2a). The three *let-7* mutations cluster in a 200-base pair (bp) segment: *let-7(mn112)* and *let-7(mg279)* are 190-bp and 27-bp deletions, respectively; and *let-7(n2853)* is a substitution within four bases of *mn112* (Fig. 2b). No protein-coding genes are predicted in the region, and no messenger RNAs were detected by probing complementary DNA libraries or by northern analysis for RNAs larger than 100 nucleotides (data not shown).

Because these studies did not identify a conventional *let-7* gene product, we used evolutionary conservation of *let-7* DNA sequences between two species of *Caenorhabditis* separated by 40 million years¹¹ to reveal probable functional sequences¹². A 2.3-kb genomic DNA fragment from *Caenorhabditis briggsae* complemented *let-7(mn112)*, showing that *let-7* function is conserved between the two species (data not shown). Comparison of *C. elegans* and *C. briggsae* *let-7* identified regions of high sequence identity (Fig. 2b). Notably, a 26-bp region flanking the *let-7(n2853)* point mutation is conserved.

We detected a 21-nucleotide RNA transcript by northern analysis

of small RNAs probed with an oligonucleotide from the conserved region flanking the *let-7* mutations (Fig. 3a). This 21-nucleotide RNA was undetectable in the *let-7(mn112)*-deletion mutant and reduced in abundance in the *let-7(n2853)* mutant. S1 nuclease analysis established the *let-7* RNA sequence as UGAGGUAGUAG-GUUGUAUAGU (Fig. 3b, c). *let-7(mn112)* deletes the first base of the transcript and 189 bases of upstream sequence, consistent with the observation that the transcript is not produced in this mutant and indicating that *let-7(mn112)* may be a null mutant. *let-7(n2853)* alters the fifth nucleotide of the *let-7* transcript. *let-7(mg279)* deletes a possible transcriptional regulatory domain upstream of the *let-7* transcript (Fig. 2b). The *let-7* RNA is unlikely to be translated: no AUG is present in any potential reading frame and the exact *C. elegans* and *C. briggsae* sequence conservation is inconsistent with the expected variation in degenerate codon positions of a translated product. The 21-nucleotide *let-7* RNA is not likely to be spliced onto a single larger transcript, because the size matches the transcript detected by northern analysis and no other transcripts were detected using the 2.5-kb rescuing fragment as a probe. One conserved region, 400bp away, contains a methionine codon (Fig. 2b), but site-directed mutagenesis of this codon in the *C. elegans* 2.5-kb rescuing fragment did not disrupt *let-7* function (data not shown). The close correspondence of the *let-7* mutations with this small RNA product, its conservation in *C. briggsae*, and the lack of open reading frames strongly support our conclusion that *let-7* functions as an RNA molecule.

let-7 expression is temporally regulated: *let-7* RNA was not detected at embryonic, L1 or L2 stages; low-level expression was

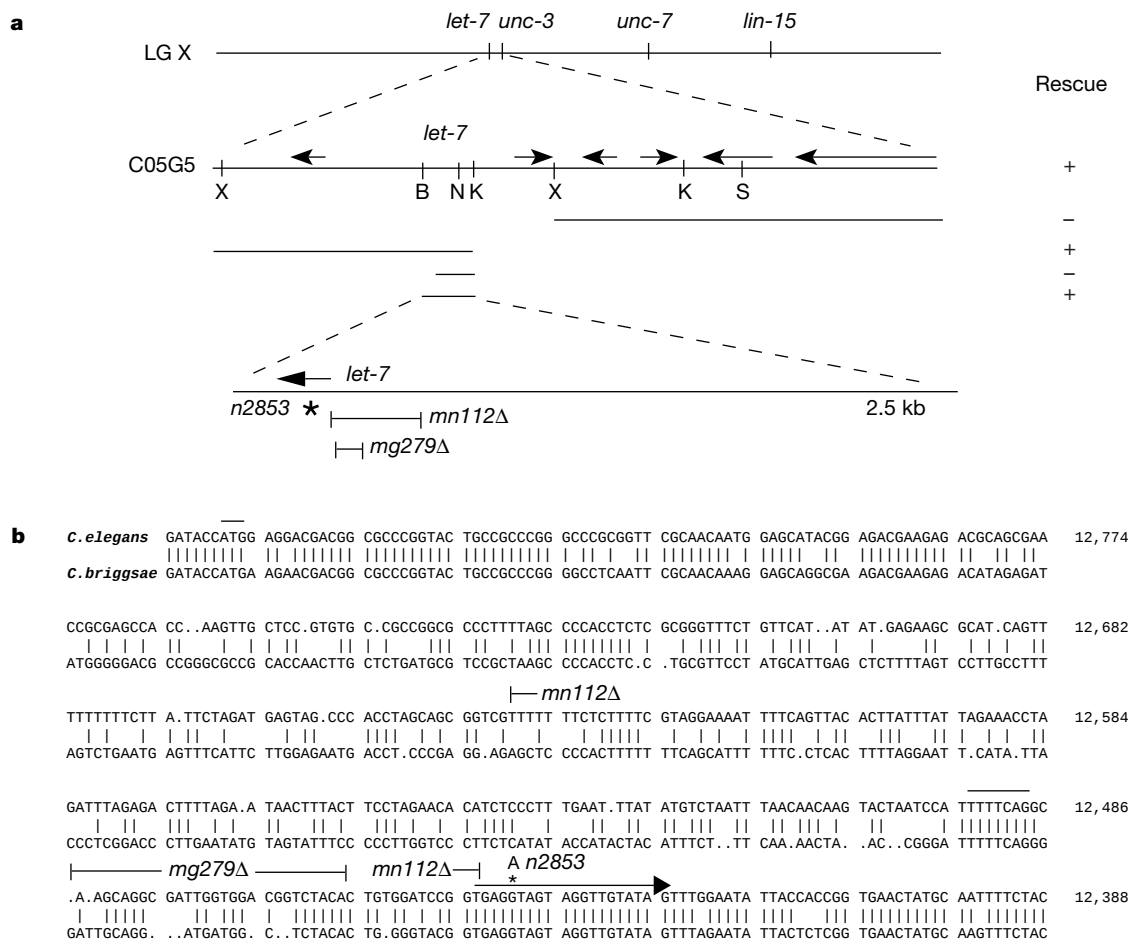


Figure 2 The *let-7* gene sequence. **a**, Transgenic rescue. +, more than 90% rescue in multiple lines; —, no rescue. Arrows above C05G5 indicate predicted genes. **b**, Sequence comparison of *let-7* from *C. elegans* and *C. briggsae*. The 2.5-kb genomic fragment that fully rescues *let-7(mn112)* corresponds to nucleotide positions 14,208–11,749 of

cosmid C05G5. The 21-nucleotide *let-7* transcript is indicated by an arrow. A truncation of the 2.5-kb fragment (to 12,446) deleting this transcript no longer rescues *let-7(mn112)*. The ATG at 12,857 and a 3' splice consensus TTTTCAG of a non-conserved open reading frame at 12,494 are indicated by bars.

detected at the early L3 stage; and high-level expression was detected at the early L4 and adult stages (Fig. 3d). This expression profile is consistent with the *let-7* mutant phenotype, which affects development specifically in late larval and adult stages. Expression of *let-7* in late larval stages also coincides with the critical period for *let-7*

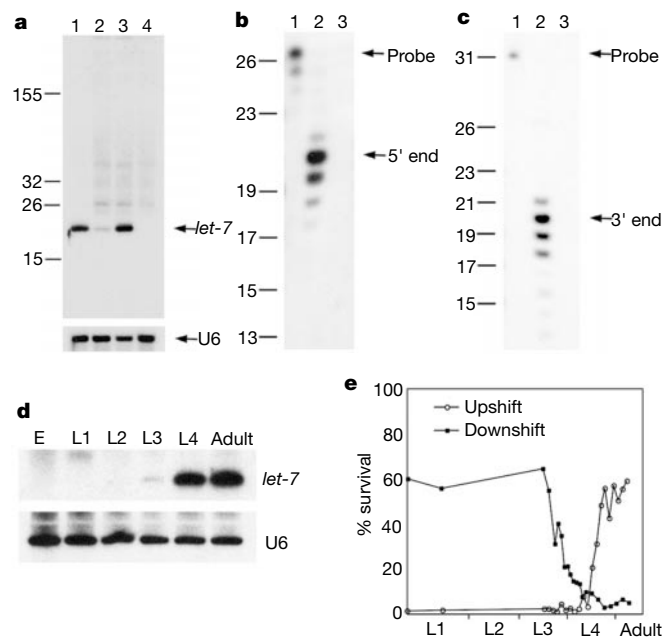


Figure 3 The 21-nucleotide *let-7* RNA. **a**, Northern blot of total RNA from mixed stage wild-type (lane 1), *let-7*(*n2853*) (lane 2), *lin-28*(*n719*) (lane 3) and *lin-28*(*n719*); *let-7*(*mn112*) *unc-3*(*e151*) animals (lane 4) probed with p249N. **b**, **c**, S1 nuclease transcript mapping. **b**, 5' probe p263 undigested (lane 1), and digested after hybridization to wild-type RNA (lane 2) or tRNA (lane 3). **c**, 3' probe p267 undigested (lane 1), and digested after hybridization to wild-type RNA (lane 2) or tRNA (lane 3). Sizing ± 1 nucleotide. **d**, Northern blot of wild-type RNA from the first 3 hours of each developmental stage. **e**, Temperature-sensitive period of *let-7*(*n2853*) viability.

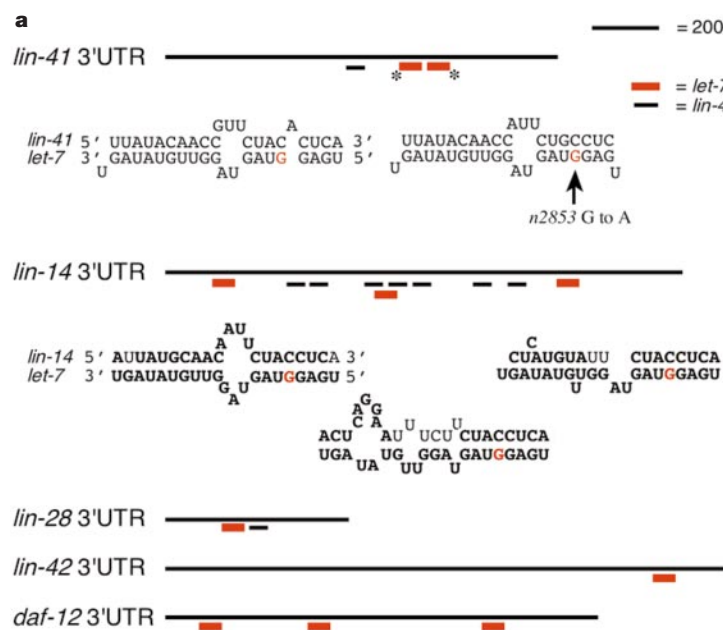


Figure 4 *let-7* regulation of heterochronic genes. **a**, *let-7* complementary sites in heterochronic genes. Bases shown in bold are conserved in genes with known *C. briggsae* homologues. **b–e**, *let-7* regulation of *lin-41*. **a**, *col-10-lacZ-lin-41* 3' UTR reporter gene was expressed in hypodermal cells of *let-7*(+) L3 larvae (**b**), downregulated in *let-7*(+)

function in viability as determined by temperature-shift experiments (Fig. 3e).

Given the genetic interactions observed between *let-7* and other heterochronic genes (Table 1) and the precedent for direct interaction between the *lin-4* regulatory RNA and genes in the heterochronic pathway, we searched for complementary regions between the *let-7* RNA and the mRNAs of these heterochronic genes. Five heterochronic genes contain sequences complementary to *let-7* in their experimentally determined (*lin-14*¹³, *lin-28*⁴ and *lin-41* (F. J. Slack *et al.*, manuscript in preparation)) or predicted (*daf-12*¹⁴ and *lin-42*¹⁵) 3' untranslated regions (UTRs) but not elsewhere in these mRNAs (Fig. 4a). The *let-7*(*n2853*) mutation would affect all predicted duplexes. The sequences of the *lin-28* element and three of the four *lin-14* elements are conserved in other Caenorhabditae.

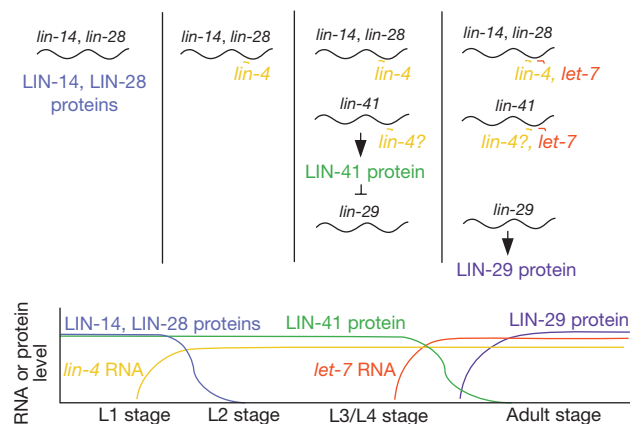


Figure 5 A model for the successive regulation of heterochronic gene activities by the *lin-4* and *let-7* RNAs. LIN-14 and LIN-28 expression levels are decreased by *lin-4* RNA expression at the end of the first larval stage to allow progression to late larval stages. In late larval stages, the expression of LIN-41 and other genes may be similarly downregulated by the *let-7* RNA, relieving their repression of LIN-29 protein expression and allowing progression to the adult stage. Because the *lin-29* mRNA does not contain sites complementary to the *let-7* RNA, *lin-29* is not likely to be a direct target of *let-7*.

The segments complementary to *let-7* are located adjacent to or overlapping segments complementary to the *lin-4* regulatory RNA in the *lin-14*, *lin-28* and *lin-41* 3' UTRs, suggesting that the two RNAs may both regulate the expression of these target genes. The *lin-29* 3' UTR lacks a *let-7* complementary region, suggesting that the regulatory interaction between *let-7* and *lin-29* is indirect. On the basis of the 36% G+C content of *C. elegans* DNA and the length of each 3' UTR, the probabilities of detecting the blocks of exact sequence complementarity to *let-7* by chance are 10^{-6} for *lin-14*, 5×10^{-5} for *lin-41*, 10^{-4} for *daf-12*, 10^{-3} for *lin-42* and 0.05 for *lin-28*. Clusters of elements were not detected in a randomly chosen set of 3' UTRs (1 isolated element in 26 3' UTRs).

Genetic analysis indicates that it may be dysregulation of *lin-41* in a *let-7* mutant that causes most of the lethal and heterochronic phenotypes (Fig. 4b–g): increasing the gene dose of *lin-41* causes the distinctive vulval bursting (Fig. 4f, g) and heterochronic phenotypes (F. J. Slack *et al.*, manuscript in preparation) of a *let-7* loss-of-function mutant; loss-of-function mutations in *lin-41* constitute the strongest class of mutations that suppress the *let-7* lethal mutant (F. J. Slack *et al.*, manuscript in preparation); and *lin-41* acts in late larval development⁹ and is downregulated at the time of *let-7* expression.

The sites that are complementary to the *let-7* RNA in the *lin-41* 3' UTR mediate *let-7*-dependent temporal downregulation of a *lacZ* reporter gene. A reporter bearing the *lin-41* 3' UTR was expressed in larval stage but not adult wild-type animals (Fig. 4b, c), similar to the temporal regulation of the *lin-41* gene itself (F. J. Slack *et al.*, manuscript in preparation), whereas a control reporter bearing the *unc-54* 3' UTR was expressed at all stages¹⁶. The *lacZ/lin-41* 3' UTR fusion gene was expressed in 79% ($n = 14$) of *let-7* (*n2853*) adult animals (Fig. 4d) but only 19% ($n = 21$) of wild-type adults. Deletion of the *let-7* complementary sites from the *lin-41* 3' UTR resulted in expression of the reporter gene in 77% of wild-type adults ($n = 30$) (Fig. 4e). These data strongly suggest that the *let-7* complementary sites in the *lin-41* 3' UTR bind to the *let-7* regulatory RNA during the L4 and adult stages to mediate downregulation of *lin-41* gene activity. The functions of *let-7* complementary sites in the 3' UTRs of *lin-42* or *daf-12* have not been assessed.

The *let-7* complementary sequences in *lin-14* and *lin-28* are more difficult to rationalize, because *lin-14* and *lin-28* function earlier in development than does *let-7* and the expression of the LIN-14 and LIN-28 proteins is downregulated by the *lin-4* regulatory RNA^{3,4,16} before the onset of *let-7* RNA expression. In fact, no major alterations in the timing or levels of LIN-14 (data not shown) or *lin-28::GFP* (green fluorescent protein)⁴ expression (V. Ambros, personal communication) were observed in *let-7* mutant animals, suggesting that either *let-7* does not regulate the expression of these genes or that *let-7* regulation of these genes is more subtle than its regulation of *lin-41*.

Two regulatory RNAs are now known to function in the heterochronic pathway. Expression of *lin-4* RNA before the second larval moult¹⁷ negatively regulates LIN-14 and LIN-28 levels to signal a transition from early to later larval patterns of cell lineage and differentiation. Expression of the *let-7* RNA during the L3 and later stages negatively regulates *lin-41* and perhaps other gene activities to signal the transition to the adult stage. Even though the *lin-4* and *let-7* RNAs are not homologous, the mechanism by which they regulate the expression of their target genes could be related. These heterochronic small RNA genes may constitute elements of a cascade of stage-specific regulatory RNAs that control the temporal sequence of events in *C. elegans* development (Fig. 5). □

Methods

Screen for retarded heterochronic genes

*egl-35(n694ts)*¹⁸ in combination with the temperature-sensitive *lin-14(n179ts)* mutation causes a sterile phenotype. To isolate suppressor mutations, we picked fertile progeny from

mutagenized *egl-35(n694ts); lin-14(n179ts)* animals. See <http://xanadu.mgh.harvard.edu/ruvkunweb/papers.html> for details concerning the genetic screen by which we identified *let-7* and the genetic mapping of *let-7*. To examine retarded heterochronic development, we observed *let-7(n2853)* animals using Nomarski optics and counted V cells and descendants at mid L1 ($n = 12$ animals), early L2 ($n = 5$), mid L2 ($n = 33$), mid L3 ($n = 29$), mid L4 ($n = 30$) and early adult ($n = 12$) stages. For the temperature-sensitive period analysis, synchronized *let-7(n2853)* animals were either upshifted or downshifted between 15° and 25° at the indicated time and scored for viability 1 day after the L4 moult.

Transgenics

For rescue experiments, genomic clones and linear PCR fragments were microinjected into *let-7* mutant animals at 2–5 ng μL^{-1} with a *goa-1::GFP* fusion gene at 70–85 ng μL^{-1} as a co-injection marker¹⁹. Transgenes with no rescuing activity (–) were maintained as balanced lines. We identified a *C. briggsae* *let-7* genomic λ clone by hybridization to a *C. elegans* *let-7* probe. A 2.3-kb *C. briggsae* region (GenBank accession number AF210771) was amplified using PCR from the λ clone and tested for rescuing activity by injection at 5 ng μL^{-1} . For *let-7* and *lin-41* overexpression experiments, either 10 ng μL^{-1} of the 2.5-kb rescuing region of *let-7* or 5 ng μL^{-1} of the *lin-41* containing cosmid C12C8 was co-injected with *goa-1::GFP* into wild-type animals. The 1148-bp *lin-41* 3' UTR was amplified using PCR and cloned 3' to the *col-10/lacZ* reporter gene¹⁶. The same transgene array that was analysed in a *let-7(+)* genetic background was crossed into the *let-7(n2853)* genetic background. To delete the *let-7* complementary sites, a deletion of 85 bp (15,565–15,650 in cosmid C12C8) in the 1148-bp *lin-41* 3' UTR region was constructed by PCR and the 1063-bp fragment was similarly cloned 3' to the *col-10/lacZ* fusion gene. Reporter genes were co-injected with a *goa-1::GFP* reporter and only GFP⁺ transgenic animals were picked for fixation and staining.

RNA analysis

RNA was isolated from *lin-28(n719); let-7(mn112) unc-3(e151)* mutants because *lin-28(n719)* suppressed the lethality of *let-7(mn112)*. Total RNA preparation, northern analysis, and S1 nuclease protection assays were done as described^{19,20}. 5' end-labelled probes were p249N, 5'-AACTATACAACTACTACTACCGGATCC-3', p263, 5'-CTATACAACTACTACTACTACCGGAT-3', pU6, 5'-GCAGGGCCATGCTAATCTTCTCTGTATTG-3'. For 3' end mapping p267, 5'-TAATATTCACAACTATA-CAACTACTACTCT-3', was annealed to p268, 5'-TGAGGTAGTAGTGTGTATAG-3', and labelled at the 3' end with Klenow and [α -³²P] dGTP. Potential RNA–RNA duplexes were identified by a combination of manual searching and computer analysis using the FOLDRNA program of the GCG software package²¹.

Received 7 October; accepted 6 December 1999.

- Ambros, V. A hierarchy of regulatory genes controls a larva-to-adult developmental switch in *C. elegans*. *Cell* **57**, 49–59 (1989).
- Ambros, V. & Horvitz, H. R. Heterochronic mutants of the nematode *Caenorhabditis elegans*. *Science* **226**, 409–416 (1984).
- Lee, R. C., Feinbaum, R. L. & Ambros, V. The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell* **75**, 843–854 (1993).
- Moss, E. G., Lee, R. C. & Ambros, V. The cold shock domain protein LIN-28 controls developmental timing in *C. elegans* and is regulated by the *lin-4* RNA. *Cell* **88**, 637–646 (1997).
- Menely, P. M. & Herman, R. K. Lethals, steriles, and deficiencies in a region of the X chromosome of *Caenorhabditis elegans*. *Genetics* **92**, 99–115 (1979).
- Sulston, J. E. & Horvitz, H. R. Post-embryonic cell lineages of the nematode *Caenorhabditis elegans*. *Dev. Biol.* **56**, 110–156 (1977).
- Rougvie, A. E. & Ambros, V. The heterochronic gene *lin-29* encodes a zinc finger protein that controls a terminal differentiation event in *Caenorhabditis elegans*. *Development* **121**, 2491–2500 (1995).
- Bettinger, J. C., Lee, K. & Rougvie, A. E. Stage-specific accumulation of the terminal differentiation factor LIN-29 during *Caenorhabditis elegans* development. *Development* **122**, 2517–2527 (1996).
- Liu, Z. *Genetic Control of Stage-Specific Developmental Events in C. elegans*. Thesis, Harvard Univ. (1990).
- Abrahante, J. E., Miller, E. A. & Rougvie, A. E. Identification of heterochronic mutants in *Caenorhabditis elegans*: temporal misexpression of a collagen::green fluorescent protein fusion gene. *Genetics* **149**, 1335–1351 (1998).
- Kennedy, B. P. *et al.* The gut esterase gene (*ges-1*) from the nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae*. *J. Mol. Biol.* **229**, 890–908 (1993).
- Zucker-Sparison, E. & Blumenthal, T. Potential regulatory elements of nematode vitellogenin genes revealed by interspecies sequence comparison. *J. Mol. Evol.* **28**, 487–496 (1989).
- Wightman, B. *et al.* Negative regulatory sequences in the *lin-14* 3'-untranslated region are necessary to generate a temporal switch during *Caenorhabditis elegans* development. *Genes. Dev.* **5**, 1813–1824 (1991).
- Yeh, W. H. *Genes Acting Late in the Signaling Pathway for Caenorhabditis elegans* Dauer Larval Development. Ph.D. Thesis, University of Missouri, Columbia, MO (1991).
- Jeon, M., Gardner, H. F., Miller, E. A., Deshler, J. & Rougvie, A. E. Similarity of the *C. elegans* developmental timing protein LIN-42 to circadian rhythm proteins. *Science* **286**, 1141–1146 (1999).
- Wightman, B., Ha, I. & Ruvkun, G. Posttranscriptional regulation of the heterochronic gene *lin-14* by *lin-4* mediates temporal pattern formation in *C. elegans*. *Cell* **75**, 855–862 (1993).
- Feinbaum, R. & Ambros, V. The timing of *lin-4* RNA accumulation controls the timing of postembryonic developmental events in *C. elegans*. *Dev. Biol.* **210**, 87–95 (1999).
- Trent, C., Tsung, N. & Horvitz, H. R. Egg-laying defective mutants of the nematode *Caenorhabditis elegans*. *Genetics* **104**, 619–647 (1983).
- Segalat, L., Elkes, D. A. & Kaplan, J. M. Modulation of serotonin-controlled behaviors by *G_o* in *Caenorhabditis elegans*. *Science* **267**, 1648–1651 (1995).
- Ausubel, F. M. *et al.* *Current Protocols in Molecular Biology* (John Wiley & Sons, New York, 1995).
- Devereux, J., Haeblerli, P. & Smithies, O. A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res.* **12**, 387–395 (1984).

Acknowledgements

We thank the *C. elegans* Genome Sequencing Consortium for sequence data, A. Coulson and the Sanger Centre for cosmids, the *Caenorhabditis* Genetics Center and V. Ambros for providing strains and sharing unpublished results. We thank R. Feinbaum for advice concerning experimental procedures and Y. Liu and P. Delorme for technical assistance. This work was supported by NIH grants to G.R., H.R.H. and A.R., and an NIH postdoctoral fellowship to E.S. H.R.H. is an Investigator of the Howard Hughes Medical Institute.

Correspondence and requests for materials should be addressed to G.R.

Pgh1 modulates sensitivity and resistance to multiple antimalarials in *Plasmodium falciparum*

Michael B. Reed^{*}, Kevin J. Saliba[†], Sonia R. Caruana^{*}, Kiaran Kirk[†] & Alan F. Cowman^{*}

^{*} The Walter and Eliza Hall Institute of Medical Research, Melbourne, Victoria 3050, Australia

[†] Division of Biochemistry and Molecular Biology, Faculty of Science, Australian National University, Canberra 0200, Australia

Throughout the latter half of this century, the development and spread of resistance to most front-line antimalarial compounds used in the prevention and treatment of the most severe form of human malaria has given cause for grave clinical concern. Polymorphisms in *pfmdr1*, the gene encoding the P-glycoprotein homologue 1 (Pgh1) protein of *Plasmodium falciparum*, have been linked to chloroquine resistance¹; Pgh1 has also been implicated in resistance to mefloquine and halofantrine^{2–5}. However, conclusive evidence of a direct causal association between *pfmdr1* and resistance to these antimalarials has remained elusive, and a single genetic cross has suggested that Pgh1 is not involved in resistance to chloroquine and mefloquine⁶. Here we provide direct proof that mutations in Pgh1 can confer resistance to mefloquine, quinine and halofantrine. The same mutations influence parasite resistance towards chloroquine in a strain-specific manner and the level of sensitivity to the structurally unrelated compound, artemisinin. This has important implications for the development and efficacy of future antimalarial agents.

Two alleles of the *pfmdr1* gene identified in field isolates of *P. falciparum* are linked with chloroquine resistance (CQR). One of these, the '7G8 allele', encodes four amino-acid substitutions with respect to the chloroquine-sensitive (CQS) 'D10 allele': Tyr 184 to Phe 184; Ser 1034 to Cys 1034; Asn 1042 to Asp 1042; and Asp 1246 to Tyr 1246 (refs 1–7). To examine the role of the last three mutations of Pgh1 in controlling parasite sensitivity and resistance to antimalarials, we constructed plasmids for *P. falciparum* transformation and allelic exchange at the endogenous *pfmdr1* locus⁸.

Plasmid pHCl1-mdr^{7G8} replaced the *pfmdr1* gene in CQS D10 parasites such that the protein carried the mutations Cys 1034, Asp 1042 and Tyr 1246. Plasmid pHCl1-mdr^{D10} (Fig. 1a) served as a transfection control and resulted in retention of the amino acids Ser 1034, Asn 1042 and Asp 1246 in Pgh1. In this manner, we generated (1) the parasite line D10-mdr^{D10} which retained the wild-type *pfmdr1* sequence, (2) the parasite line D10-mdr^{7G8/3} into which the *pfmdr1* gene encoding the Cys 1034, Asp 1042 and Tyr 1246 mutations was inserted, and (3) the parasite line D10-mdr^{7G8/1} which encoded the Tyr 1246 mutation in *pfmdr1* owing to a single recombination event in the gene between the codons encoding this amino acid and position 1042 (Fig. 1a). Analysis of genomic DNA

by Southern hybridization (Fig. 1b) and sequencing of the *pfmdr1* gene confirmed these integration events.

To determine the role of the Cys 1034, Asp 1042 and Tyr 1246 substitutions in a distinct genetic background, we made similar constructs for CQR 7G8 parasites and carried out analogous experiments. Plasmid pHH1-mdr^{D10} (Fig. 1c) allowed allelic replacement of the *pfmdr1* gene within 7G8 parasites such that the gene encoded the wild-type (D10) amino acids Ser 1034, Asn 1042 and Asp 1246. The two cloned lines, 7G8-mdr^{D10/c1} and 7G8-mdr^{D10/c2}, were generated in this manner (Fig. 1c). pHH1-mdr^{7G8} (Fig. 1c) served as a transfection control and, once integrated, retained the mutant *pfmdr1* allele (7G8-mdr^{7G8} parasites; Fig. 1c).

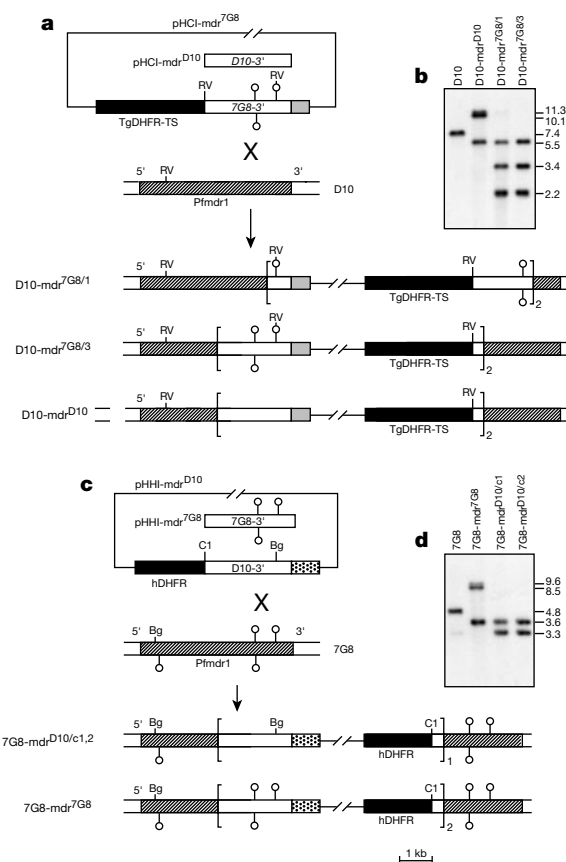


Figure 1 Allelic replacement of the *pfmdr1* gene. **a**, Allelic replacement of the *pfmdr1* gene in the D10 cloned parasite line. The transfection plasmids pHCl1-mdr^{7G8} and pHCl1-mdr^{D10} are shown. Open circles indicate the mutations Cys 1034, Asp 1042 and Tyr 1246 in pHCl1-mdr^{7G8} (ref. 1). The codon for Tyr 1246 creates an *EcoRV* site that was used to map the integration events for this plasmid. The selection cassette *Tgdhfr-ts* (*Toxoplasma gondii* dihydrofolate reductase-thymidylate synthase), which confers resistance to pyrimethamine^{20–22} is indicated. The integration structure for D10-mdr^{7G8/1}, in which the recombination event occurred between the Asp 1042 and Tyr 1246 polymorphisms in the pHCl1-mdr^{7G8} plasmid resulting in the introduction of only the Tyr 1246 mutation in the endogenous *pfmdr1* gene, and the structures of the plasmid integration events in D10-mdr^{7G8/3} (for plasmid pHCl1-mdr^{7G8}) and D10-mdr^{D10} (for plasmid pHCl1-mdr^{D10}) are shown. All integration events occurred through a single recombination event resulting in reconstitution of the *pfmdr1* gene and displacement of a fragment of the gene downstream with insertion of two copies of the plasmid in each case⁸. RV, *EcoRV*. **b**, Southern hybridization of genomic DNA digested with *EcoRV* from each parasite line. **c**, Allelic replacement of the *pfmdr1* gene in the 7G8 cloned parasite line. The transfection plasmids pHH1-mdr^{D10} and pHH1-mdr^{7G8} are shown. The selection cassette includes the human *dhfr* gene. Integration events are shown for the two clones 7G8-mdr^{D10/c1,2} and 7G8-mdr^{7G8}. The codon for Asp 1246 creates a *Bgl* II site. Bg, *Bgl* II; Cl, *Clal*. **d**, Southern hybridization of *Bgl* II/*Clal*-digested genomic DNA from each parasite line. Size of DNA fragments are shown in kb (**b,d**).

Southern blot analysis (Fig. 1d) and sequencing of the *pfmdr1* gene verified the integration and relevant gene sequence. Immunoblot analysis using anti-Pgh1 antibodies⁹ showed that each transfected line expressed equivalent levels of Pgh1.

To test the role of Cys 1034, Asp 1042 and Tyr 1246 in determining drug resistance, we carried out *in vitro* drug assays¹⁰ for chloroquine, mefloquine, quinine, halofantrine and artemisinin (Fig. 2; and Table 1). Introduction of the *pfmdr1* polymorphisms into CQS D10 parasites had no effect on parasite sensitivity to chloroquine (Fig. 2; and Table 1). In contrast, replacement of the 7G8 mutations with wild-type D10 *pfmdr1* sequence in CQR 7G8 parasites halved the level of CQR (Fig. 2; and Table 1). These results show that *pfmdr1* polymorphisms are insufficient to confer CQR in D10. In 7G8, however, *pfmdr1* provides a cumulative effect with additional gene(s) to confer higher levels of CQR. This indicates that *pfmdr1* mutations in *P. falciparum* may have been selected by chloroquine pressure and may be required for resistance to high levels of drug.

Introduction of the 7G8 mutations into the D10 *pfmdr1* gene (D10-mdr^{7G8/3}) converted a quinine-sensitive isolate into one that is quinine resistant. We confirmed this by subsequent removal of the

same mutations from 7G8 (7G8-mdr^{D10/c1} and 7G8-mdr^{D10/c2}), which resulted in reversion to sensitivity for quinine (Fig. 2; and Table 1). These results indicate that mutations in *pfmdr1* alone are sufficient to confer quinine resistance in the D10 and 7G8 genetic backgrounds. Notably, the 7G8 mutations can confer quinine resistance in D10 and yet have no effect on CQR in this genetic background. It is possible, however, that quinine resistance is also dependent on multiple genes and that *pfmdr1* contributes to the overall phenotype.

The *pfmdr1* substitutions had a marked effect on susceptibility to mefloquine and halofantrine. Introduction of the Cys 1034, Asp 1042 and Tyr 1246 into D10 Pgh1 conferred increased sensitivity to both drugs (Fig 2; and Table 1). Conversely, altering *pfmdr1* within 7G8 (mefloquine- and halofantrine-sensitive) parasites to encode Ser 1034, Asn 1042 and Asp 1246 conferred resistance against both antimalarials. Notably, the single Tyr 1246 mutation (D10-mdr^{7G8/1}) had a larger effect on the half-maximal inhibitory concentration (IC₅₀) for both mefloquine and halofantrine and suggests that this amino-acid position is directly involved in mefloquine accumulation (see below). Clearly, the presence of the Cys 1034, Asp 1042 and Tyr 1246 *pfmdr1* substitutions in the field would provide a disadvantage in terms of a parasite's ability to survive the use of mefloquine and/or halofantrine in these areas. The converse would be true for quinine and, in some parasite backgrounds, chloroquine.

Insertion of the three mutations into D10 *pfmdr1* increased the level of artemisinin sensitivity almost twofold (Fig. 2; and Table 1). Conversely, removal of the mutations from 7G8 caused a decrease in sensitivity by about the same margin. The *pfmdr1* effect on artemisinin susceptibility mimics the results seen with mefloquine and halofantrine. This is consistent with a common mechanism whereby *pfmdr1* can influence the accumulation of mefloquine, halofantrine and artemisinin, a hypothesis that is consistent with field studies that show a positive correlation between the IC₅₀ for artemisinin and that of mefloquine and halofantrine¹¹.

Removal of the Cys 1034, Asp 1042 and Tyr 1246 mutations from 7G8 *pfmdr1* increased saturable, steady-state accumulation of chloroquine in 7G8-mdr^{D10/c1} and 7G8-mdr^{D10/c2} as compared with 7G8-mdr^{7G8} ($P = 0.01$ and $P = 0.035$, respectively; Fig. 3a), which is consistent with the reduction in the chloroquine IC₅₀ (Fig. 2; and Table 1). A linear inverse correlation between the saturable accumulation of chloroquine and the IC₅₀ for inhibition of parasite growth has been previously demonstrated¹². The twofold increase in accumulation by 7G8-mdr^{D10/c1} and 7G8-mdr^{D10/c2} relative to 7G8-mdr^{7G8} (Fig. 3a) therefore accounts for the twofold decrease in the IC₅₀ of chloroquine. No change in chloroquine accumulation was observed for either D10-mdr^{7G8/1} or D10-mdr^{7G8/3}. In contrast, the

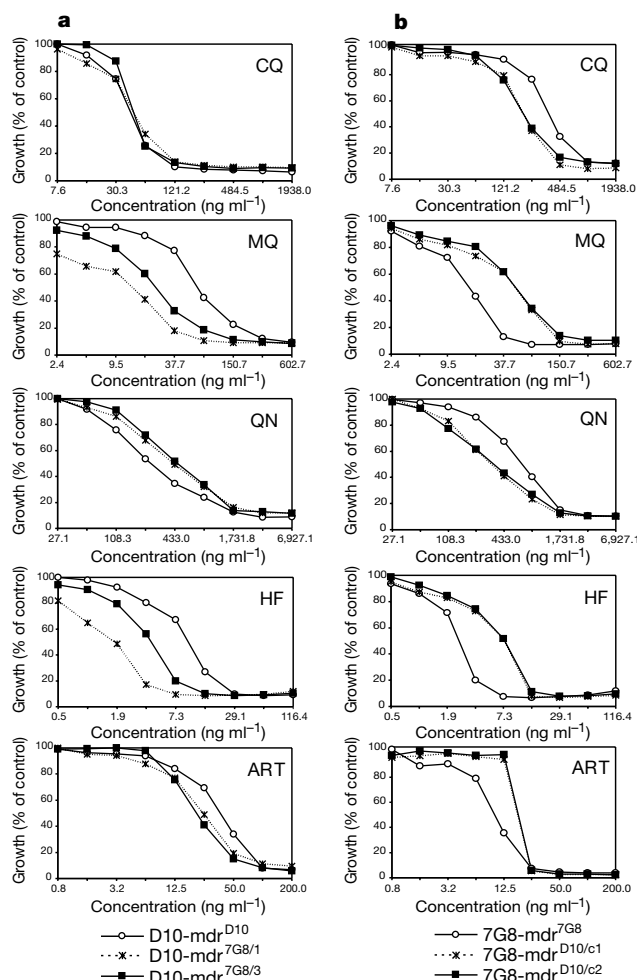


Figure 2 Effect of introduction/removal of the polymorphisms in the D10 and 7G8 cloned lines on susceptibility and resistance to chloroquine, mefloquine, halofantrine, quinine and artemisinin. **a**, Microdilution assays to determine drug susceptibility of D10 compared with the transfectants D10-mdr^{D10}, D10-mdr^{7G8/1} and D10-mdr^{7G8/3}. The drugs used are indicated: CQ, chloroquine; MQ, mefloquine; QN, quinine; HF, halofantrine; ART, artemisinin. Experiments shown were performed twice in triplicate. **b**, As in panel **a**, but microdilution assays were also carried out using the parasite lines 7G8, 7G8-mdr^{7G8}, 7G8-mdr^{D10/c1}, 7G8-mdr^{D10/c2}.

Table 1 Effect of mutations in the *pfmdr1* gene on the sensitivity and resistance to diverse antimalarials

Parasite	50% Inhibitory concentration, IC ₅₀ (nM)*				
	Chloroquine	Mefloquine	Quinine	Halofantrine	Artemisinin
D10	47.9 ± 1.6	73.3 ± 6.5	269.5 ± 18.0	9.5 ± 0.6	42.2 ± 5.3
D10-mdr ^{D10}	45.2 ± 3.7	71.1 ± 4.6	272.6 ± 33.5	10.4 ± 0.6	38.7 ± 4.6
D10-mdr ^{7G8/1}	52.7 ± 4.8	14.9 ± 1.4†	413.0 ± 53.2	1.7 ± 0.2†	23.8 ± 2.1†
D10-mdr ^{7G8/3}	48.4 ± 0.8	25.8 ± 1.9†	505.0 ± 83.9†	4.3 ± 0.4†	21.6 ± 1.4†
7G8	382.4 ± 8.9	15.2 ± 0.2	585.6 ± 18.6	2.1 ± 0.2	10.9 ± 0.6
7G8-mdr ^{7G8}	389.5 ± 12.4	16.9 ± 1.7	692.5 ± 62.0	2.6 ± 0.2	12.3 ± 1.5
7G8-mdr ^{D10/c1}	204.1 ± 1.2‡	55.2 ± 2.9‡	310.0 ± 29.1‡	7.6 ± 0.4‡	18.6 ± 0.5‡
7G8-mdr ^{D10/c2}	215.5 ± 15.3‡	53.3 ± 7.0‡	378.7 ± 66.8‡	7.6 ± 1.1‡	19.0 ± 0.4‡
Threshold of resistance†	100	25–30	450–500	5–20	–

* Values represent the mean (± s.e.m.) of two independent assays each performed in triplicate.
† The threshold of resistance corresponds to values previously defined^{11,27} and have been obtained by determining serum drug concentration in patients infected with sensitive and resistant *P. falciparum*.
‡ When compared with the relevant controls the changes in IC₅₀ values were significant as assessed using the Mann–Whitney U test ($P < 0.05$).

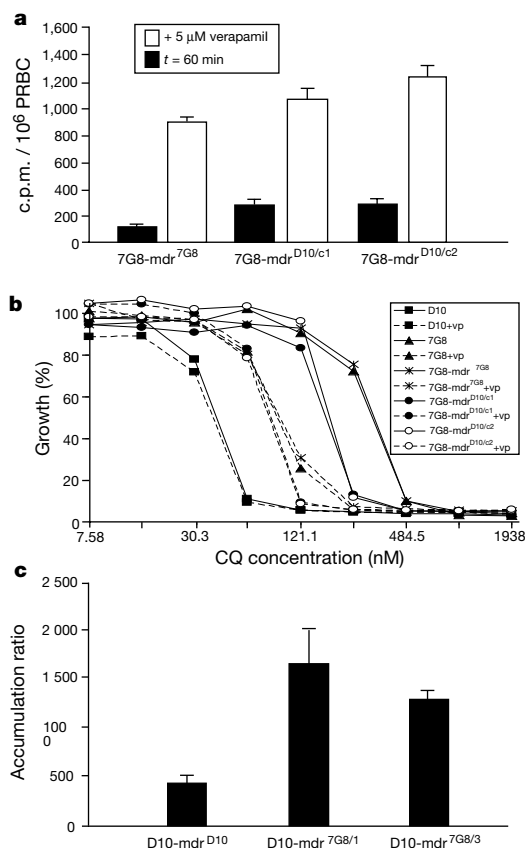


Figure 3 Effect of *pfmdr1* mutations on accumulation of antimalarials and modulation with verapamil. **a**, Saturable chloroquine accumulation and ability of verapamil (vp; 5 μM) to increase chloroquine accumulation in 7G8-mdr^{7G8}, 7G8-mdr^{D10/c1} and 7G8-mdr^{D10/c2}. PRBC, parasitized red blood cells. The accumulation was measured over 60 min. Experiments shown were performed three times, each in duplicate. Error bars, s.e.m. **b**, Microdilution assays to determine chloroquine sensitivity in the presence and absence of 5 μM verapamil for D10, 7G8, 7G8-mdr^{7G8}, 7G8-mdr^{D10/c1} and 7G8-mdr^{D10/c2}. **c**, Mefloquine cellular accumulation ratios for D10-mdr^{D10}, D10-mdr^{7G8/1} and D10-mdr^{7G8/3} as estimated by the inoculum effect. Two independent experiments were each carried out in duplicate. Error bars, s.e.m. The difference in accumulation between D10-mdr^{D10} and the transfected lines D10-mdr^{7G8/1} and D10-mdr^{7G8/3} were significantly different ($P < 0.05$).

change in phenotype from mefloquine resistance to mefloquine sensitivity was reflected in increased accumulation of mefloquine for both D10-mdr^{7G8/1} and D10-mdr^{7G8/3} relative to D10-mdr^{D10} ($P < 0.05$; Fig. 3c)¹³.

Drug resistance and accumulation of chloroquine in *P. falciparum* can be reversed by verapamil¹⁴, although the level of chloroquine sensitivity obtained is never equivalent to that seen for CQS lines¹⁵. This indicates that a component of CQR may remain verapamil insensitive and supports the hypothesis that the CQR phenotype is multigenic¹⁶. In the presence of verapamil, the chloroquine accumulation (Fig. 3a) and chloroquine sensitivity (Fig. 3b) of 7G8-mdr^{7G8}, 7G8-mdr^{D10/c1} and 7G8-mdr^{D10/c2} were the same. Addition of verapamil caused a sevenfold increase in the saturable accumulation of chloroquine by 7G8-mdr^{7G8}, and a fourfold increase in the saturable accumulation by both 7G8-mdr^{D10/c1} and 7G8-mdr^{D10/c2}, which is consistent with the observation that verapamil caused a larger shift in the chloroquine IC₅₀ for 7G8-mdr^{7G8} than for 7G8-mdr^{D10/c1} and 7G8-mdr^{D10/c2} (Fig. 3a, b). Therefore, the increased CQR attributable to *pfmdr1* occurs within the component of the phenotype sensitive to verapamil, though it remains to be established whether Pgh1 is a target or if reversal occurs by an alternative mechanism¹⁷.

After a decade of fierce debate, these data provide the first direct evidence, to our knowledge, of the involvement of Pgh1 in conferring high levels of CQR through decreased chloroquine accumulation. Nevertheless, it is clear that CQR cannot be conferred by Pgh1 alone and requires the presence of mutations in other (unidentified) gene(s)^{1,6,18}. In addition, we have shown that changes in Pgh1 can modulate resistance to quinine, mefloquine and halofantrine. The Cys 1034, Asp 1042 and Tyr 1246 mutations in *pfmdr1* are distributed in Africa, South America and Asia, which suggests that there is selective pressure for their maintenance and spread¹. This is consistent with an important role for these *pfmdr1* mutations in antimalarial drug resistance throughout these geographical areas. That these mutations can confer resistance and sensitivity in two cloned lines originating from Papua New Guinea (D10 parasites) and South America (7G8 parasites) is further support for this view. The finding that there are strains of *P. falciparum* in malaria-endemic areas showing decreased sensitivity to artemisinin has serious implications for the future prospects of this important antimalarial.

Methods

Plasmid construction

pSP72-based plasmids (Promega) bearing the complete coding regions of *pfmdr1* (ref. 7) cloned as single *XhoI* fragments from D10 (Papua New Guinea) and 7G8 (South America) *P. falciparum* isolates (plasmids designated p12 and p12-7G8, respectively) were digested with *EcoRI/XhoI* to release 2.8-kilobase (kb) 3'-terminal fragments of *pfmdr1*. After 'fill-in' reactions using Vent polymerase (New England Biolabs), the 2.8-kb DNA fragments were subcloned into *SmaI*-digested pBluescript-SK (Stratagene). Plasmids bearing inserts in the required orientation were subsequently chosen to allow release of the complete *pfmdr1* insert as a *XhoI* fragment. To generate plasmids pHC1-mdr^{D10} and pHC1-mdr^{7G8}, these *XhoI* inserts derived from D10 and 7G8 parasites were then cloned into the *P. falciparum* transfection vector pHC1 (ref. 19). To generate a transfection vector suitable for transforming pyrimethamine-resistant parasites (including 7G8), the *Toxoplasma gondii* *dhfr-ts* region contained within pHC4 (refs 20–22) was replaced by a human *dhfr* fragment mutated to encode resistance to methotrexate and WR99210 (ref. 23). This human *dhfr* fragment was derived by polymerase chain reaction (PCR) from pH22Y. The resultant vector pHH1 was then digested with *HincII/XhoI* to remove the *hsp86* upstream regulatory sequences and enable cloning of the 2.8-kb *pfmdr1* fragments. The resultant plasmid constructs were designated pHH1-mdr^{D10} and pHH1-mdr^{7G8}.

Parasite transformation

Transformation of plasmids pHC1-mdr^{D10} and pHC1-mdr^{7G8} into D10 parasites and pHH1-mdr^{D10}, pHH1-mdr^{7G8} into 7G8 parasites was carried out using the low-voltage, high-capacitance conditions described²³. Selection for pyrimethamine resistance (pHC1-based plasmids) was carried out essentially as described⁸, and selection for stable transfection and integration of pHH1-based vectors was carried out in the same manner using 5 nM WR99210 (kindly provided by D. Jacobus). Where appropriate, parasites were cloned by limiting dilution²⁴.

IC₅₀ determinations

Sensitivity to antimalarials was assessed using a modification of the standard microdilution technique described¹⁰. For this modification, predominantly ring-stage parasites (1% parasitaemia, 4% haematocrit) were plated out in hypoxanthine-free media in the presence of [8-³H]hypoxanthine (Amersham) for the duration of the assay. Drug sensitivity was plotted as the percentage of growth relative to growth in drug-free medium and IC₅₀ values were determined graphically. Each assay was performed in triplicate on two independent occasions.

Chloroquine accumulation

Chloroquine accumulation in cells suspended in bicarbonate-free RPMI medium, at a final haematocrit of 2% and parasitaemia of between 1.3–6.4% was measured as described¹⁵. [³H]chloroquine (1 nM and 50.4 Ci mmol⁻¹) was incubated with cells for 60 min at 37 °C before separating the cells from the suspending medium by centrifugation through an oil layer¹⁵. The cell pellets were lysed with 0.5% v/v Triton X-100, bleached with 1% w/v sodium hypochlorite, then neutralized with HCl before scintillation counting. Chloroquine is accumulated by parasitized erythrocytes in both saturable and non-saturable processes, and it is the former process that has been correlated with the antiparasitic action of the drug¹². The saturable component of [³H]chloroquine accumulation was estimated by subtracting the non-saturable accumulation, measured in the presence of 10 μM unlabelled chloroquine, from the total accumulation, measured in the absence of unlabelled drug. The small contribution of uninfected cells to the uptake of [³H]chloroquine in parasitized cell suspensions was subtracted in all cases.

Mefloquine accumulation

Because of the lipophilic nature of mefloquine, previous investigations have indicated a significant 'inoculum effect' with this drug^{2,3}. As a result, measured drug IC₅₀ increases with increasing inoculum size (inoculum size = parasitaemia × haematocrit) owing to significant drug depletion from the medium. This phenomenon has been described in detail^{13,15,26}. In this study, IC₅₀ of mefloquine was assessed at inoculum sizes ranging from 1 to 10 (fractional parasite volume 0.0001 to 0.001) at a fixed haematocrit of 4%. Extrapolation of the linear relationship between measured IC₅₀ and inoculum size provides a measure of absolute drug IC₅₀ at a theoretical inoculum of zero. The mathematical relationship for the determination of the cellular drug accumulation ratio is

$$\text{Accumulation ratio} = \frac{\text{IC}_{50} \text{ measured} - \text{IC}_{50} \text{ absolute}}{\text{IC}_{50} \text{ absolute} \times \text{fractional parasite volume}}$$

Parasites were synchronized with sorbitol 48 h before accumulation studies and two independent assays were performed, each in duplicate.

Received 10 September; accepted 21 December 1999.

1. Foote, S. J. *et al.* Several alleles of the multidrug-resistance gene are closely linked to chloroquine resistance in *Plasmodium falciparum*. *Nature* **345**, 255–258 (1990).
2. Wilson, C. M. *et al.* Amplification of a gene related to mammalian mdr genes in drug-resistant *Plasmodium falciparum*. *Science* **244**, 1184–1186.
3. Peel, S. A., Merritt, S. C., Handy, J. & Baric, R. S. Derivation of highly mefloquine-resistant lines from *Plasmodium falciparum* *in vitro*. *Am. J. Trop. Med. Hyg.* **48**, 385–397.
4. Cowman, A. F., Galatis, D. & Thompson, K. J. Selection for mefloquine resistance in *Plasmodium falciparum* is linked to amplification of the *pfdmr1* gene and cross-resistance to halofantrine and quinine. *Proc. Natl Acad. Sci. USA* **91**, 1143–1147 (1994).
5. Peel, S. A., Bright, P., Yount, B., Handy, J. & Baric, R. S. A strong association between mefloquine and halofantrine resistance and amplification, overexpression, and mutation in the P-glycoprotein gene homolog (*pfdmr*) of *Plasmodium falciparum* *in vitro*. *Am. J. Trop. Med. Hyg.* **51**, 648–658 (1994).
6. Welles, T. E. *et al.* Chloroquine resistance not linked to mdr-like genes in a *Plasmodium falciparum* cross. *Nature* **345**, 253–255 (1990).
7. Foote, S. J., Thompson, J. K., Cowman, A. F. & Kemp, D. J. Amplification of the multidrug resistance gene in some chloroquine-resistant isolates of *P. falciparum*. *Cell* **57**, 921–930 (1989).
8. Triglia, T., Wang, P., Sims, P. F. G., Hyde, J. E. & Cowman, A. F. Allelic exchange at the endogenous genomic locus in *Plasmodium falciparum* proves the role of dihydropteroate synthase in sulfadoxine-resistant malaria. *EMBO J.* **17**, 3807–3815 (1998).
9. Cowman, A. F., Karcz, S., Galatis, D. & Culvenor, J. G. A P-glycoprotein homologue of *Plasmodium falciparum* is localized on the digestive vacuole. *J. Cell Biol.* **113**, 1033–1042 (1991).
10. Desjardins, R. E., Canfield, C. J., Haynes, J. D. & Chulay, J. D. Quantitative assessment of antimalarial activity *in vitro* by a semiautomated microdilution technique. *Antimicrob. Agents Chemother.* **16**, 710–718 (1979).
11. Doury, J. C., Ringwald, P., Guclain, J. & Le Bras, J. Susceptibility of African isolates of *Plasmodium falciparum* to artemisinin (qinghaosu). *Trop. Med. Parasitol.* **43**, 197–198 (1992).
12. Bray, P. G., Munghin, M., Ridley, R. G. & Ward, S. A. Access to hematin: the basis of chloroquine resistance. *Mol. Pharmacol.* **54**, 170–179 (1998).
13. Hawley, S. R., Bray, P. G., O'Neill, P. M., Park, B. K. & Ward, S. A. The role of drug accumulation in 4-aminoquinoline antimalarial potency. The influence of structural substitution and physicochemical properties. *Biochem. Pharmacol.* **52**, 723–733 (1996).
14. Martin, S. K., Oduola, A. M. & Milhous, W. K. Reversal of chloroquine resistance in *Plasmodium falciparum* by verapamil. *Science* **235**, 899–901 (1987).
15. Bray, P. G., Hawley, S. R., Munghin, M. & Ward, S. A. Physicochemical properties correlated with drug resistance and the reversal of drug resistance in *Plasmodium falciparum*. *Mol. Pharmacol.* **50**, 1559–1566 (1996).
16. Ward, S. A., Bray, P. G., Munghin, M. & Hawley, S. R. Current views on the mechanisms of resistance to quinoline-containing drugs in *Plasmodium falciparum*. *Ann. Trop. Med. Parasitol.* **89**, 121–124 (1995).
17. Bray, P. G. *et al.* Cellular uptake of chloroquine is dependent of binding to ferriprotoporphyrin IX and is independent of NHE activity in *Plasmodium falciparum*. *J. Cell Biol.* **145**, 363–376 (1999).
18. Su, X. -Z., Kirkman, L. A., Fujioka, H. & Welles, T. E. Complex polymorphisms in an ~330 kDa protein are linked to chloroquine-resistant *P. falciparum* in Southeast Asia and Africa. *Cell* **91**, 593–603 (1997).
19. Crabb, B. S., Triglia, T., Waterkeyn, J. G. & Cowman, A. F. Stable transgene expression in *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **90**, 131–144 (1997).
20. Wu, Y., Kirkman, L. A. & Welles, T. E. Transformation of *Plasmodium falciparum* malaria parasites by homologous integration of plasmids that confer resistance to pyrimethamine. *Proc. Natl Acad. Sci. USA* **93**, 1130–1134 (1996).
21. Crabb, B. S. & Cowman, A. F. Characterization of promoters and stable transfection by homologous and nonhomologous recombination in *Plasmodium falciparum*. *Proc. Natl Acad. Sci. USA* **93**, 7289–7294 (1996).
22. Crabb, B. S. *et al.* Targeted gene disruption shows that knobs enable malaria-infected red cells to cytoadhere under physiological shear stress. *Cell* **89**, 287–296 (1997).
23. Fidock, D. A. & Welles, T. E. Transformation with human dihydrofolate reductase renders malaria parasites insensitive to WR99210 but does not affect the intrinsic activity of proguanil. *Proc. Natl Acad. Sci. USA* **94**, 10931–10936 (1997).
24. Kirkman, L. A., Su, X. Z. & Welles, T. E. *Plasmodium falciparum*: isolation of large numbers of parasite clones from infected blood samples. *Exp. Parasitol.* **83**, 143–149 (1996).
25. Ritchie, G. Y. *et al.* *In vitro* selection of halofantrine resistance in *Plasmodium falciparum* is not associated with increased expression of Pgh1. *Mol. Biochem. Parasitol.* **83**, 35–46 (1996).
26. Geary, T. G., Divo, A. D., Jensen, J. B., Zangwill, M. & Ginsburg, H. Kinetic modelling of the response of *Plasmodium falciparum* to chloroquine and its experimental testing *in vitro*. Implications for mechanism of action and of resistance to the drug. *Biochem. Pharmacol.* **40**, 685–691 (1990).
27. Cremer, G., Basco, L. K., Le Bras, J. L., Camus, D. & Slomianny, C. *Plasmodium falciparum*: detection of P-glycoprotein in chloroquine-susceptible and chloroquine-resistant clones and isolates. *Exp. Parasitol.* **81**, 1–8 (1995).

Acknowledgements

We thank S. Ward and P. Bray for the gift of [³H]chloroquine and for advice throughout this work, and T. Welles and D. Fidock for providing the human *dhfr* gene. This work was supported by the National Health and Medical Research Council of Australia (NHMRC) and the National Institute of Health, USA. M.B.R. is supported by a NHMRC Peter Doherty Research Fellowship. We are grateful to the Blood Banks of the Red Cross of Victoria and Australian Capital Territory for generously providing serum and red cells.

Correspondence and requests for materials should be addressed to A.F.C. (e-mail: cowman@wehi.edu.au).

The LIM homeobox gene *Lhx9* is essential for mouse gonad formation

Ohad S. Birk*, Delane E. Casiano*, Christopher A. Wassif†, Tiziana Cogliati‡, Liping Zhao§, Yangu Zhao*, Alexander Grinberg*, SingPing Huang*, Jordan A. Kreidberg||, Keith L. Parker§, Forbes D. Porter† & Heiner Westphal*

* Laboratory of Mammalian Genes and Development and † Heritable Disorders Branch, National Institute of Child Health and Human Development, NIH, Bethesda, Maryland 20892, USA

‡ Genetics Department, Medicine Branch, National Cancer Institute, NIH, Bethesda, Maryland 20889, USA

§ Department of Internal Medicine, University of Texas Southwestern Medical Center, Dallas, Texas 75235, USA

|| Department of Medicine, Children's Hospital, Boston, Massachusetts 02115, USA

During mammalian embryonic development, the ovaries and testes develop from somatic cells of the urogenital ridges as indifferent gonads, harbouring primordial germ cells that have migrated there. After sex determination of the gonads, the testes produce testosterone and anti-Müllerian hormone which mediate male sexual differentiation, and the female developmental pathway ensues in their absence^{1–3}. Here we show that transcripts of the LIM homeobox gene *Lhx9* are present in urogenital ridges of mice at embryonic day 9.5; later they localize to the interstitial region as morphological differentiation occurs. In mice lacking *Lhx9* function, germ cells migrate normally, but somatic cells of the genital ridge fail to proliferate and a discrete gonad fails to form. In the absence of testosterone and anti-Müllerian hormone, genetically male mice are phenotypically female. The expression of steroidogenic factor 1 (Sf1), a nuclear receptor essential for gonadogenesis², is reduced to minimal levels in the *Lhx9*-deficient genital ridge, indicating that *Lhx9* may lie upstream of Sf1 in a developmental cascade. Unlike mice lacking other genes that mediate early stages of gonadogenesis^{4–6}, *Lhx9* mutants do not exhibit additional major developmental defects. Thus, *LHX9* mutations may underlie certain forms of isolated gonadal agenesis in humans.

In mouse embryos, the sexually undifferentiated gonads emerge from the urogenital ridges as distinct structures by embryonic day 12.0 (E12.0)(ref. 2). During this indifferent stage, the superficial coelomic mesothelium of the genital ridge proliferates into the subjacent loose connective mesenchymal tissue. In males (Fig. 1a), this proliferating epithelium subsequently forms sex cords that extend into the mesenchyme of the presumptive gonad. The testicular cords contain Sertoli cells which surround the primordial germ cells that have migrated to the gonadal ridge. Eventually, the cords lose contact with the surface epithelium and are separated from it by the subjacent mesenchyme that later differentiates to form the tunica albuginea. Meanwhile, the deeper layers of mesenchyme form the interstitial mesenchymal cells of the testes,

some of which differentiate into Leydig cells that make testosterone. In females (Fig. 1b), the epithelium produces cortical sex cords that do not penetrate deeply into the mesenchyme, but split into clusters that will differentiate into granulosa cells. The mesenchymal cells of the ovary differentiate into theca cells. The theca and granulosa cells together form the follicles that envelop the germ cells and secrete steroid hormones^{7,8}.

Several autosomal genes encoding transcription factors mediate early events in the development of the indifferent gonads. *Lhx1* (alias *Lim1*), Wilms tumour 1 (*Wt1*) and steroidogenic factor 1 (*Sf1*) all are expressed in the urogenital ridge by E9.5 (refs 5, 6, 9, 10). In mice with disruptions of any of these genes, gonadal development is arrested before sexual differentiation can occur, and the mice die *in utero* or shortly after birth from developmental defects in the adrenal cortex (*Sf1*)⁶, kidneys (*Wt1*)⁵, or kidneys and brain (*Lhx1*)⁴.

Sex determination leading to testes formation is initiated by *Sry*, which is encoded on the Y chromosome^{3,11,12}. Other genes implicated in regulating the male pathway of sexual differentiation include *Sf1*, *Wt1* and *Sox9* (refs 3, 11, 12). In contrast, *Dax1* expression increases in the ovary as female sexual differentiation occurs¹³. This finding has led to the speculation that *Dax1* contributes specifically to ovarian development^{13,14}, although analyses of *Dax1* knockout mice indicate that it does not have an essential role in this process¹⁵. Thus, out of the genes whose products are

known to mediate key events in sexually differentiated gonadal function after E12.5, only *Sf1* and *Wt1* are currently known to also be essential for the earlier stages of gonadal development in both sexes.

We mapped *Lhx9* (refs 16, 17), a member of the LIM homeobox domain gene family of transcription factors¹⁸, to mouse chromosome 1, and analysed its expression in the gonadal region. In both sexes, *Lhx9* was expressed in the medial surface of the E9.5 urogenital ridge in the presumptive gonadal region (Fig. 1c). By E11.5 (Fig. 1d), *Lhx9* expression was high in the mesothelial lining (the 'germinal' epithelium) and its subjacent mesenchyme (the future male tunica albuginea) and lower in the mesenchymal cells in deeper layers of the gonadal ridge. At E13.5 (Fig. 1e–g), *Lhx9* expression was high in the mesothelial layer and the outer part of the testis (the developing tunica), lower in the interstitial mesenchyme, and nearly absent in the sex cords. Thus, the E13.5 expression in the testis is compatible with *Lhx9* being downregulated in the mesothelial cells once they have formed the sex cords. In the female, the expression pattern of *Lhx9* at E9.5 and E11.5 resembles that of the male. In the E13.5 ovary (Fig. 1h–j), *Lhx9* expression was mostly limited to the cortical region.

To test whether *Lhx9* is involved in gonadal development, we deleted exons 2 and 3 of the *Lhx9* gene encoding the first and second LIM domains by homologous recombination in murine embryonic stem (ES) cells (see Methods). Mice heterozygous for the *Lhx9* mutation appeared normal and were fertile. In crosses between these heterozygotes, homozygous mutant offspring were born at the expected frequency (53 out of 207, 26%) and were fully viable. Although polymerase chain reaction (PCR) analysis for the Y chromosome genes *Sry* and *Zfy*¹⁹ showed that half of these homo-

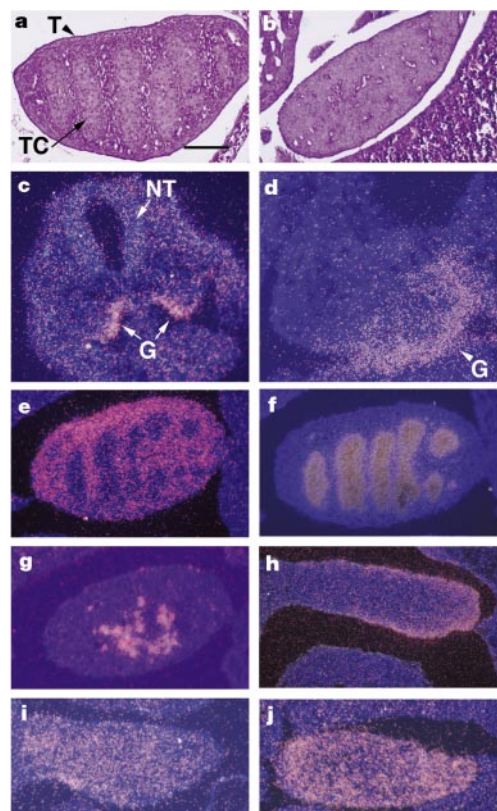


Figure 1 *Lhx9* expression in the developing murine gonad. **a, b**, Haematoxylin and eosin staining of E13.5 testis (**a**) and ovary (**b**). **c, d**, *Lhx9* transcripts in the developing gonadal ridge as detected by *in situ* hybridization of transverse sections of wild-type E9.5 embryo (**c**) and E11.5 urogenital ridge (**d**). *Lhx9* localizes specifically to the gonadal component of the ridge (**d**). **e–g**, Serial sections of E13.5 testis. The expression pattern of *Lhx9* (**e**) differs from those of *Amh* (specific to Sertoli cells) (**f**) and *P450sc* (**g**), a marker identifying Leydig cells²². **h–j**, The expression pattern of *Lhx9* in the E13.5 ovary (**h**) differs from those of *Dax1* (**i**) and *Sf1* (**j**). Scale bar shown in **a** represents 80 μ m in **a, b, e–j** and 50 μ m in **c, d**. T, developing tunica; TC, testicular cords; G, gonadal ridge; NT, neural tube.

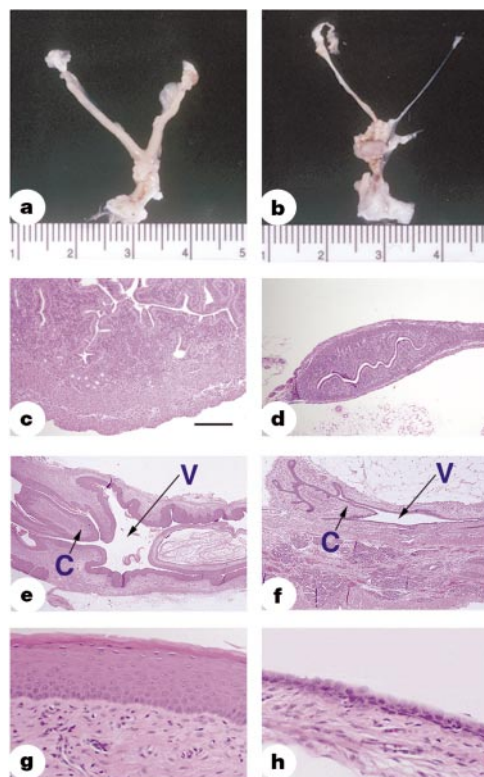


Figure 2 Reproductive tract of 3-month-old *Lhx9*^{+/+} female (**a, c, e, g**) and sibling *Lhx9*^{-/-} mutant male (**b, d, f, h**). **a, b**, Macroscopic view of the uterus. **c, d**, Haematoxylin and eosin staining of a cross section through the uterine horn. **e, f**, Mid-sagittal section through the cervix (C) and vagina (V). **g, h**, Epithelial lining of the vagina. Scale bar in **c** represents 200 μ m for **c, d**, 1,000 μ m for **e, f** and 100 μ m for **g, h**. The mutant phenotype was identical in three lines originating from three different founder mice.

zygous *Lhx9* mutant mice were genetically male (not shown), all mice were phenotypically female and had no gonads. Consistent with the absence of gonads, both genetically male and female *Lhx9*^{-/-} mice were infertile, had atrophic uteri, vaginas and oviducts (Fig. 2), had no male accessory sex organs or oestrus cycle, and had markedly elevated levels of follicle-stimulating hormone with undetectable levels of testosterone and oestradiol (data not shown).

Although *Lhx9* is normally expressed in the embryonic central nervous system, limbs and pancreas^{16,17}, we found no gross abnormalities in these structures in *Lhx9*^{-/-} mice. This probably reflects a redundancy of *Lhx9* and other LIM homeobox genes with overlapping expression patterns, such as its close structural relative, *Lhx2* (refs 16, 17). Nevertheless, further detailed analysis of these structures in the *Lhx9* mutants is warranted.

To explore the basis for the gonadal agenesis in the *Lhx9* mutant mice, we analysed the structures of the developing gonads at different stages of embryogenesis. At E11.5, *Lhx9*^{-/-} and wild-type urogenital ridges were morphologically indistinguishable (Fig. 3a, b). By E12.0, clear budding was evident in the wild-type genital ridge (Fig. 3c), but no such proliferation was seen in the *Lhx9*^{-/-} mutants (Fig. 3d). By E13.5, when the wild-type ovaries and testes had clearly emerged as discrete structures (Fig. 3e, g), no gonads were visible in the *Lhx9*^{-/-} embryos (Fig. 3f, h).

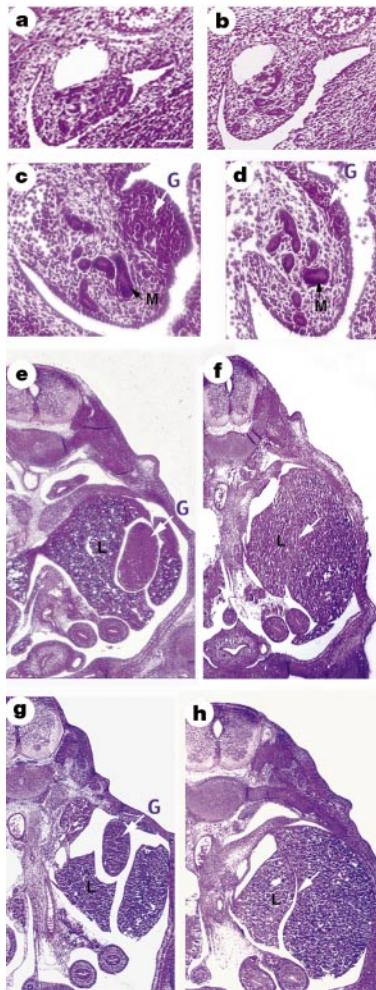


Figure 3 Arrest of gonad formation in *Lhx9* mutants. Haematoxylin and eosin staining of transverse sections through *Lhx9*^{+/+} (a, c, e, g) and *Lhx9*^{-/-} (b, d, f, h) sibling embryos at E11.5 (a, b), E12.0 (c, d) and E13.5 (female (e, f); male (g, h)). Arrows point to the gonad (G) or to its destined location (from which it is absent) in the mutants. L, liver; M, mesonephric tubules. Scale bar in a represents 80 μm for a, b, 60 μm for c, d and 300 μm for e-h.

A critical event in gonadal development is the migration of the primordial germ cells into the indifferent gonad. Alkaline phosphatase staining showed that these primordial germ cells migrated into the presumptive *Lhx9*^{-/-} gonadal region at the expected time (Fig. 4a, b). Unlike the gonads of *Wt1* and *Sf1* knockout mice, which exhibit histological evidence of apoptosis at about E11.5–E12.5^{5,6}, the gonads of *Lhx9*^{-/-} embryos lacked evidence for DNA fragmentation (Fig. 4c, d). However, the massive cell proliferation normally seen in the genital ridge at E11.5 (Fig. 4e) was markedly attenuated in the mutant embryos (Fig. 4f). We did not detect transcripts of the non-disrupted 5' end of *Lhx9* in E11.5 *Lhx9*^{-/-} gonadal ridges (Fig. 4g, h), suggesting that *Lhx9* is necessary for the proliferation of gonadal cells in which it is normally transcribed. Thus, the proliferative effect of *Lhx9* seems to be, at least in part, cell autonomous. In contrast to the findings in the *Lhx9*^{-/-} gonadal ridge, the E11.5 *Lhx9*^{-/-} brain and limbs exhibited BrdU incorporation and expression of the non-disrupted 5' end of *Lhx9* that were indistinguishable from the *Lhx9*^{+/+} controls (data not shown). This is probably due to redundancy of *Lhx9* with other proteins, such as *Lhx2*, which is known to be expressed both in the embryonic limbs and brain but not in the gonadal ridge^{17,20}. Alternatively, it is possible that the proliferative effect of *Lhx9* is specific to the gonads, and occurs only when *Lhx9* is expressed in conjunction with other molecules that are present in the gonadal ridge but not at other sites of *Lhx9* expression. At any rate, the failure in gonad formation in the *Lhx9* mutants reflects a failure of cell proliferation rather than the initiation of programmed cell death.

To gain insights into possible interactions between *Lhx9* and other genes involved in early gonadogenesis (for example, *Lhx1*, *Wt1*, *Sf1*, *Dax1*), we examined their expression in *Lhx9* mutant embryos. In E11.5 *Lhx9*^{-/-} embryos, *Sf1* transcripts were markedly reduced in the gonadal ridge but not in the adjacent adrenal bud (Fig. 4i, j). This suggests that *Sf1* is differentially regulated in various tissues. Some *Sf1* transcripts were still present in the E11.5 *Lhx9*^{-/-} gonadal ridge, which suggests that *Lhx9* may not regulate *Sf1* directly, but, instead, permit the proliferation of *Sf1*-positive cells. The effect of *Lhx9* on *Sf1* seen at E11.5 might well be cell autonomous, as both transcripts exhibit partly overlapping expression patterns in the gonadal ridge at this time (Fig. 4g, i). The diminished expression of *Sf1* in the E11.5 *Lhx9*^{-/-} urogenital ridge, coupled with the uninterrupted expression of *Lhx9* in *Sf1*^{-/-} embryos (Fig. 4k, l), indicates that *Lhx9* may act upstream of *Sf1* in a developmental cascade leading to the formation of the mammalian gonads. Comparing the E11.5 *Lhx9*^{-/-} and *Lhx9*^{+/+} gonadal ridges, we observed no change in the expression pattern of *Wt1* (Fig. 4m, n). This is in line with the normal expression pattern of *Wt1* in the E11.5 *Sf1*^{-/-} gonadal ridge². Thus, the data imply that in the E11.5 gonadal ridge, *Sf1* marks a cell population different from *Wt1*, and that the expression of *Wt1* is independent of *Lhx9* and *Sf1*. *In situ* hybridization demonstrated normal localization of *Lim1* in the *Lhx9*^{-/-} gonadal ridge at E9.5 and *Dax1* at E11.5 (data not shown), but a change of dosage cannot be ruled out. By E13.0, the *Lhx9*^{-/-} gonads had regressed completely, precluding analyses of gene expression.

We have shown that the autosomal gene *Lhx9* is expressed in the epithelial and subjacent mesenchymal cells of the early gonadal ridge, where it permits cell proliferation and *Sf1* expression, and is essential in the formation of the indifferent gonad. The morphology (Fig. 3d) and the minimal *Sf1* expression (Fig. 4j) seen in the E11.5 *Lhx9*^{-/-} presumptive gonadal ridge imply that some differentiation does occur there, indicating that *Lhx9* may not direct the initial commitment to a gonadal fate. The expression pattern of *Lhx9*, the *Lhx9*^{-/-} phenotype and the BrdU data together suggest that signals controlled by *Lhx9* are responsible both for the growth of the gonads and for the proliferation and invasion of the epithelium into the mesenchyme—an event essential for formation of the sex cords. There is ample evidence that LIM homeobox genes control

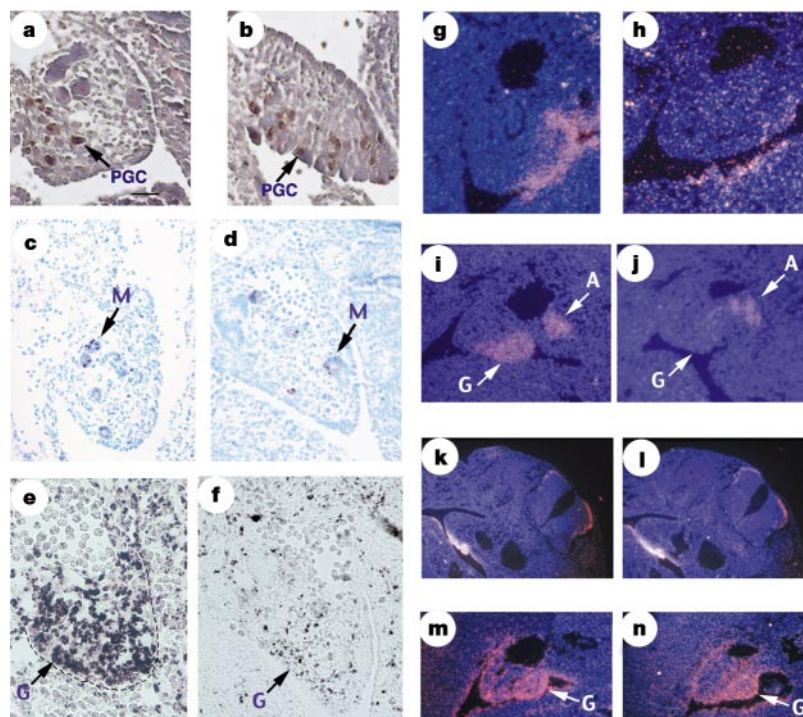


Figure 4 Analysis of the defects in gonad formation in *Lhx9* mutant embryos (**b, d, f, h, j, n**) as compared with wild-type controls (**a, c, e, g, i, m**). **a, b**, Alkaline phosphatase staining demonstrating primordial germ cells (PGC) in the urogenital ridge of E12.0 embryos. **c, d**, TUNEL staining of the urogenital ridge at E11.5. Note the apoptosis in mesonephric tubules (M) in both wild-type and *Lhx9*^{-/-} sections. **e, f**, Proliferating cells in the gonadal ridge (G) at E10.5, as labelled by anti-BrdU staining. Dotted line delineates the urogenital

ridge. **g, h**, *Lhx9* transcripts 5' to the disrupted region, in E11.5 gonadal ridge. **i, j**, *Sfl* transcripts in the gonadal ridge (G) and the adrenal primordium (A) at E11.5. **k, l**, *Lhx9* transcripts in the gonadal ridge of *Sfl*^{+/+} (**k**) and *Sfl*^{-/-} (**l**) E11.5 embryos. **m, n**, *Wt1* transcripts in the gonadal ridge (G) at E11.5. **a–n** are transverse sections. Scale bar in **a** represents 60 µm for **a–f** and 100 µm for **g–n**.

proliferation in distinct embryonic cell types¹⁸ (for example, *Lhx2* in the developing forebrain²⁰). Our data indicate that *Lhx9* may have an analogous role in the developing gonad. Human studies and analysis of null mutant mice implicate other genes such as *Sfl*, *Wt1* and *Lim1* in early gonadal development. However, null mutations in these genes also cause other major defects^{4–6}, whereas a null mutation in the murine *Lhx9* gene specifically affects gonadal development. Failure of gonad formation in the absence of other syndromic features has been well documented in human beings²¹. It is therefore possible that mutations in *LHX9* might underlie some forms of isolated human gonadal agenesis. □

Methods

Generation of *Lhx9*^{-/-} mice

The targeting vector was constructed by replacing 2 kilobases (kb) containing exons 2 and 3 of the *Lhx9* gene with a PGK-NEO cassette, preserving 4.5-kb (5') and 2.2-kb (3') flanks of homologous sequences and using the diphtheria toxin A (DTA) gene for double selection²³. Homologous recombination in murine ES cells was achieved²³, and confirmed by Southern blot analysis (details available upon request). Chimaeric mice, generated as described²³, were mated with wild-type C57BL/6 mice to generate *Lhx9*^{+/+} animals that were crossed to produce offspring for analysis.

Immunohistochemistry and *in situ* hybridization

Primordial germ cells were detected using alkaline phosphatase staining⁶. Cells undergoing apoptotic cell death were detected by the terminal deoxynucleotidyl transferase (TdT)-mediated dUTP nick end labelling (TUNEL) method using Trevigen TACS-2 kit²³. To label embryonic proliferating cells, pregnant females were injected intraperitoneally at E10.5 with BrdU (100 µg per gram of body weight) and killed 2 hours later. Immunohistochemical detection of BrdU-labelled cells was done as described²³. RNA probes for *Lhx9* (ref. 16), *Lhx1* (ref. 9), *Sfl* (ref. 6), *Wt1* (ref. 5), *Dax1* (ref. 13), *Amh* (ref. 3) and *P450sc* (ref. 22) were labelled with ³³P using Ambion MAXIScript kit. *In situ* hybridization was done as described²³.

Polymerase chain reaction

Lhx9 wild-type allele was detected using primers GAATGTGAAATGGGGGAA and TAAGATGCCACTGTTTGTCTACGG, *Lhx9*^{-/-} allele with primers CTTGGGTGGAGAG GCTATTC and AGGTGAGATGACAGGAGATC, *Sry* and *Zfy* as described¹⁹.

Chromosomal mapping of mouse *Lhx9*

Primers for detection of *Lhx9* wild-type allele (above) and the T31 radiation hybrid panel RH04.02 (Research Genetics) were used according to manufacturer's instructions.

Received 2 November 1999; accepted 17 January 2000.

- Jost, A. Problems of fetal endocrinology: the gonadal and hypophyseal hormones. *Rec. Prog. Horm. Res.* **8**, 379–418 (1953).
- Parker, K. L., Schedl, A. & Schimmer, B. P. Gene interactions in gonadal development. *Annu. Rev. Physiol.* **61**, 417–433 (1999).
- Swain, A. & Lovell-Badge, R. Mammalian sex determination: a molecular drama. *Genes Dev.* **13**, 755–767 (1999).
- Shawlot, W. & Behringer, R. R. Requirement for *Lim1* in head organizer function. *Nature* **374**, 425–430 (1995).
- Kreidberg, J. A. *et al.* WT-1 is required for early kidney development. *Cell* **74**, 679–691 (1993).
- Luo, X., Ikeda, Y. & Parker, K. L. A cell-specific nuclear receptor is essential for adrenal and gonadal development and sexual differentiation. *Cell* **77**, 481–490 (1994).
- Kaufman, M. H. & Bard, J. B. L. in *The Anatomical Basis of Mouse Development* 117–128 (Academic, New York, 1999).
- Gilbert, S. F. *Developmental Biology* 5th edn (Sinauer Associates, Sunderland, MA, 1997).
- Fujii, T. *et al.* Expression patterns of the murine LIM class homeobox gene *lim1* in the developing brain and excretory system. *Dev. Dyn.* **199**, 73–83 (1994).
- Barnes, J. D., Crosby, J. L., Jones, C. M., Wright, C. V. & Hogan, B. L. Embryonic expression of *Lim-1*, the mouse homolog of *Xenopus* *Xlim-1*, suggests a role in lateral mesoderm differentiation and neurogenesis. *Dev. Biol.* **161**, 168–178 (1994).
- Capel, B. Sex in the 90s: SRY and the switch to the male pathway. *Annu. Rev. Physiol.* **60**, 497–523 (1998).
- Koopman, P. Sry and Sox9: mammalian testis-determining genes. *Cell. Mol. Life Sci.* **55**, 839–856 (1999).
- Swain, A. *et al.* Mouse *Dax1* expression is consistent with a role in sex determination as well as in adrenal and hypothalamus function. *Nature Genet.* **12**, 404–409 (1996).
- Goodfellow, P. N. & Camerino, G. *Dax-1*, an 'antitestis' gene. *Cell. Mol. Life Sci.* **55**, 857–863 (1999).
- Yu, R. N., Masafumi, I., Saunders, T. L., Camper, S. A. & Jameson, J. L. Role of *Ahch* in gonadal development and gametogenesis. *Nature Genet.* **20**, 353–357 (1998).
- Bertuzzi, S. *et al.* Characterization of *lhx9*, a novel LIM/homeobox gene expressed by the pioneer neurons in the mouse cerebral cortex. *Mech. Dev.* **81**, 193–198 (1999).
- Sylvie Retaux, S. *et al.* *Lhx9*, a novel LIM-homeodomain gene expressed in the developing forebrain. *J. Neurosci.* **19**, 783–793 (1999).
- Hobert, O. & Westphal, H. Functions of LIM-homeobox genes. *Trends Genet.* **16**, 75–83 (2000).
- Kunieda, T. *et al.* Sexing of mouse preimplantation embryos by detection of Y chromosome-specific sequences using polymerase chain reaction. *Biol. Reprod.* **46**, 692–697 (1992).
- Porter, F. D. *et al.* *Lhx2*, a LIM homeobox gene, is required for eye, forebrain and definitive erythrocyte development. *Development* **124**, 2935–2944 (1997).

21. Warne, G. L. & Zajac, J. D. Disorders of sexual differentiation. *Endocrinol. Metab. Clin. North Am.* **27**, 945–967 (1998).
22. Ikeda, Y. *et al.* Developmental expression of mouse steroidogenic factor 1, an essential regulator of the steroid hydroxylases. *Mol. Endocrinol.* **8**, 654–662 (1994).
23. Zhao, Y. *et al.* Control of hippocampal morphogenesis and neuronal differentiation by the LIM homeobox gene *Lhx5*. *Science* **284**, 1155–1158 (1999).

Acknowledgements

We thank R. Lovell-Badge for *Dax1*, *Amh* and *Sox9* probes; J.B.L. Bard, K. Pfeifer and A. Tomac for comments on the manuscript; L. Nieman for fruitful discussions; and J. Levine for assistance. L.Z. and K.L.P. are supported by the NIH.

Correspondence and requests for materials should be addressed to H.W. (e-mail: hw@helix.nih.gov).

The S receptor kinase determines self-incompatibility in *Brassica* stigma

Takeshi Takasaki[†], Katsunori Hatakeyama^{*}, Go Suzuki[‡], Masao Watanabe[§], Akira Isogai^{||} & Kokichi Hinata^{*}

^{*} Research Institute of Seed Production Co., Ltd. 6-6-3 Minamiyoshinari, Aoba-ku, Sendai 989-3204, Japan

[†] Faculty of Agriculture, Kobe University, Nada-ku, Kobe 657-8501, Japan

[‡] Division of Natural Science, Osaka Kyoiku University, Kashiwara, Osaka 582-8582, Japan

[§] Faculty of Agriculture, Iwate University, Morioka 020-8550, Japan

^{||} Graduate School of Biological Sciences, Nara Institute of Science and Technology, Ikoma 630-0101, Japan

The self-incompatibility possessed by *Brassica* is an intraspecific reproductive barrier by which the stigma rejects self-pollen but accepts non-self-pollen for fertilization. The molecular/biochemical bases of recognition and rejection have been intensively studied. Self-incompatibility in *Brassica* is sporophytically controlled by the polymorphic *S* locus¹. Two tightly linked polymorphic genes at the *S* locus, *S* receptor kinase gene (*SRK*) and *S* locus glycoprotein gene (*SLG*), are specifically expressed in the papillar cells of the stigma^{2–4}, and analyses of self-compatible lines^{5–7} of *Brassica* have suggested that together they control stigma function in self-incompatibility interactions. Here we show, by transforming self-incompatible plants of *Brassica rapa* with an *SRK*²⁸ and an *SLG*²⁸ transgene separately, that expression of *SRK*²⁸ alone, but not *SLG*²⁸ alone, conferred the ability to reject self (*S*²⁸)-pollen on the transgenic plants. We also show that the ability of *SRK*²⁸ to reject *S*²⁸ pollen was enhanced by *SLG*²⁸. We conclude that *SRK* alone determines *S* haplotype specificity of the stigma, and that *SLG* acts to promote a full manifestation of the self-incompatibility response.

S receptor kinase is a membrane-spanning receptor kinase that consists of an extracellular domain (called the *S* domain), a transmembrane domain and a cytosolic domain with serine/threonine kinase activity⁸. *S* locus glycoprotein is a secreted glycoprotein^{3,4} whose amino-acid sequence is highly similar, but not identical, to the *S* domain of *SRK* (ref. 2). It has been proposed that *SRK* and *SLG* function together as the receptor for a pollen ligand, which determines the *S* specificity of pollen, and that this receptor–ligand interaction sets off a cascade of biochemical reactions leading to the self-incompatibility response^{9,10}. A cysteine-rich gene located at the *S* locus, called *SP11* (ref. 11) or *SCR*¹², has been shown by gain-of-function and loss-of-function experiments to encode the potential pollen ligand¹². However, the proposed role of

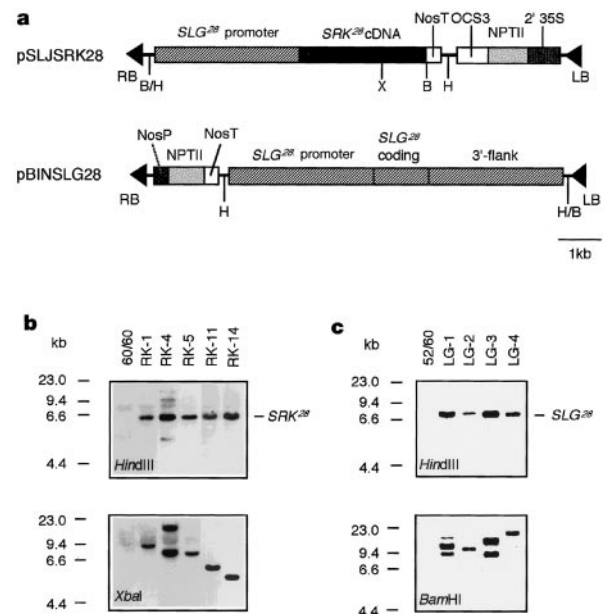


Figure 1 Vector construction and detection of transgenes. **a**, Representation of the T-DNA region of transformation vectors pSLJSRK28 and pBINSLG28. H, *HindIII*; X, *XbaI*; B, *BamHI*; NosT, *nos* terminator; OCS3, the 3' region of *ocs*; 2', *mas2'* promoter; 35S, cauliflower mosaic virus 35S promoter; NosP, *nos* promoter. RB, right border; LB, left border. **b**, Representative DNA gel blot analysis of the presence (upper panel) and its copy numbers (lower panel) of the *SRK*²⁸ transgene in an *S*⁶⁰ homozygote (60/60) and five independent transgenic plants (RK-1, RK-4, RK-5, RK-11 and RK-14) using the cDNA encoding the kinase domain of *SRK*²⁸ (*SRK*²⁸-KD) as a probe. **c**, DNA gel blot analysis of the presence (upper panel) and its copy numbers (lower panel) of the *SLG*²⁸ transgene in an *S*⁵²*S*⁶⁰ heterozygote (52/60) and four independent transgenic plants (LG-1, LG-2, LG-3 and LG-4) using *SLG*²⁸ cDNA as a probe.

SRK and *SLG* in self-incompatibility interactions has so far been based entirely on circumstantial evidence^{2–7}. The most direct way to show the function of *SRK* or *SLG* would be to transform *Brassica* plants with the *SLG* or *SRK* gene of a different *S* haplotype and show that the transgenic plants acquire the *S* haplotype specificity of the transgene. However, all transformation experiments reported so far have resulted in the breakdown of self-incompatibility in the transgenic plants because of co-suppression between the endogenous *SLG* and/or *SRK* gene and the *SLG* and/or *SRK* transgene^{13–15}.

We chose to examine the function of *SRK* by introducing an *SRK* gene of a class I *S* haplotype, *SRK*²⁸, into plants homozygous for a class II *S* haplotype, *S*⁶⁰. This classification of *S* haplotypes is based on amino-acid sequence similarities between their *SLG* proteins and between their *SRK* proteins¹⁶. An *S* haplotype can be dominant over, recessive to, or co-dominant with another *S* haplotype, but class II *S* haplotypes are, in most cases, recessive to class I *S* haplotypes in pollen¹⁷. We considered that the lower degree of nucleotide sequence similarity between *SRK* genes of different classes of *S* haplotypes than between *SRK* genes of the same class might minimize the problem of co-suppression encountered in all the previous transformation experiments. We constructed a transformation vector, pSLJSRK28 (Fig. 1a), and introduced it into *S*⁶⁰ homozygotes of *B. rapa* by *Agrobacterium*-mediated transformation. To examine the function of *SLG*, we constructed another transformation vector, pBINSLG28 (Fig. 1a), and introduced it into *S*⁵²*S*⁶⁰ heterozygotes of *B. rapa* by the same transformation method (*S*⁵² is a class I *S* haplotype). The *SLG*²⁸ promoter used in both constructs had previously been shown to be active in the stigma but not in the anther^{18,19}.

In the case of pSLJSRK28 transformation, the *SRK*²⁸ transgene was detected in 17 independent transgenic plants by DNA gel blot

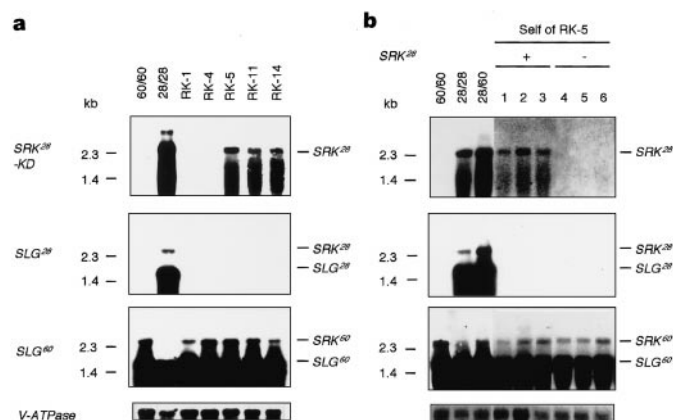


Figure 2 RNA blot analysis of transcription of the endogenous *SRK²⁸* gene and the *SRK²⁸* transgene. The probes used are shown to the left of the blots. **a**, Transcription of *SRK²⁸* in an *S⁶⁰* homozygote (60/60) and an *S²⁸* homozygote (28/28), and five primary transgenic plants (RK-1, RK-4, RK-5, RK-11 and RK-14). **b**, Transcription of *SRK²⁸* in three control plants, *S⁶⁰* homozygote (60/60), *S²⁸* homozygote (28/28) and *S²⁸S⁶⁰* heterozygote (28/60), and six plants of the self-pollinated progeny (self) of RK-5. The presence (+) or absence (–) of the *SRK²⁸* transgene is indicated above the blots.

analysis (Fig. 1b). All of them were self-incompatible; three (RK-5, RK-11 and RK-14) rejected both *S²⁸* and *S⁶⁰* pollen, whereas the other 14 plants rejected only *S⁶⁰* pollen. The pollen of the former three plants was rejected by *S⁶⁰* but not by *S²⁸* stigmas. These results suggest that the self-incompatibility phenotype of the stigmas of RK-5, RK-11 and RK-14 had changed from that of the parental *S⁶⁰* homozygotes, but the self-incompatibility phenotype of the pollen had not. Further, the self-incompatibility phenotype of the other 14 transgenic plants was not altered.

To determine whether the new self-incompatibility phenotype of the stigmas of RK-5, RK-11 and RK-14 resulted from the expression of the *SRK²⁸* transgene, we first pollinated these three plants with pollen from *S²⁴*, *S⁴³*, *S⁴⁵* and *S⁵²* homozygotes (all being class I *S* haplotypes), and *S²⁹* and *S⁴⁴* homozygotes (class II *S* haplotypes) of *B. rapa*¹⁷. In all pollinations, pollen tubes penetrated the stigma, indicating that the new self-incompatibility phenotype was specific to the *S²⁸* haplotype. We then carried out RNA gel blot analysis and detected the *SRK²⁸* transcript in the stigmas of these three plants, but not in the stigmas of the transgenic plants that did not acquire the ability to reject *S²⁸* pollen (Fig. 2a). The expression level of *SRK²⁸* in these three plants averaged 32–35% of that in *S²⁸* heterozygote. The endogenous *SRK⁶⁰* and *SLG⁶⁰* were expressed at normal levels in all the transgenic plants, and, as expected, the *SRK²⁸* transcript was not

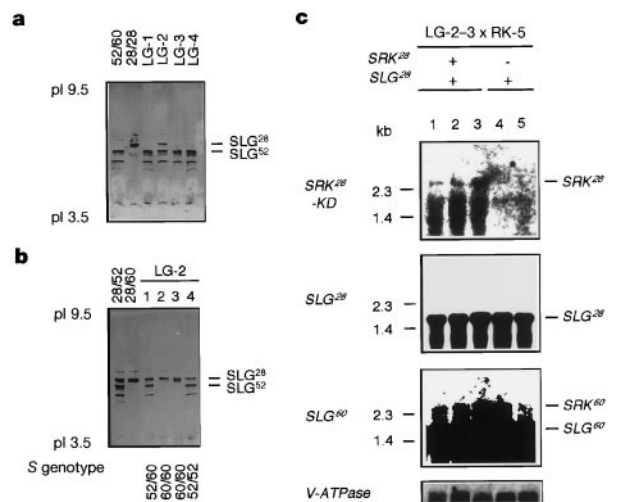


Figure 3 Expression of the *SRK²⁸* and *SLG²⁸* transgenes. **a**, Isoelectric focusing immunoblot analysis of the expression of the endogenous *SLG²⁸* gene in an *S⁵²S⁶⁰* heterozygote (52/60) and an *S²⁸* homozygote (28/28), and the expression of the *SLG²⁸* transgene in four primary transgenic plants (LG-1, LG-2, LG-3 and LG-4). **b**, Isoelectric focusing immunoblot analysis of the expression of the endogenous *SLG²⁸* gene in an *S²⁸S⁵²* heterozygote (28/52) and an *S²⁸S⁶⁰* heterozygote (28/60), and the expression of the *SLG²⁸* transgene in four of the plants (1–4) obtained from bud self-pollination of LG-2. The genotypes of the latter four plants are shown below the blot. **c**, Transcription of the *SRK²⁸* and *SLG²⁸* transgenes in the LG-2-3 × RT-5 progeny. The presence (+) or absence (–) of the *SRK²⁸* and *SLG²⁸* transgenes is indicated above the blots. The probes are indicated to the right of the blots.

detected in the anther of any of these transgenic plants (data not shown).

We examined whether the *S²⁸* haplotype specificity acquired by the stigmas of RK-5 co-segregated with the *SRK²⁸* transgene in the progeny. As this plant was self-incompatible (with the stigma and pollen phenotypes being *S²⁸S⁶⁰* and *S⁶⁰*, respectively), we carried out self-pollination at immature bud stages when self-incompatibility was not manifested. Polymerase chain reaction (PCR) analysis showed that 16 out of the 20 plants in the progeny inherited the *SRK²⁸* transgene and DNA gel blot analysis (Fig. 1b) indicated that RK-5 carried a single copy of the *SRK²⁸*. Thus, the segregation ratio of the *SRK²⁸* transgene in the self-pollinated progeny of RK-5 was consistent with mendelian inheritance of a single copy of the transgene ($\chi^2 = 0.1$, $P > 0.7$, degrees of freedom (d.f.) = 1). As was the case for RK-5, the stigmas of all the 16 progeny plants that carried the *SRK²⁸* transgene rejected both *S²⁸* and *S⁶⁰* pollen, and

Table 1 Self-incompatibility (SI) phenotypes of plants in self-pollinated and crossed progeny of RK-5

Progeny	S genotype	SI phenotype		No. of plants	No. of flowers	Seeds/flower
		Stigma	Pollen			
Selfing of RK-5	<i>S⁶⁰/S⁶⁰, SRK²⁸/SRK²⁸ or –/–</i>	<i>S²⁸S⁶⁰</i>	<i>S⁶⁰</i>	16		
	<i>S⁶⁰/S⁶⁰, –/–</i>	<i>S⁶⁰</i>	<i>S⁶⁰</i>	4		
<i>S⁶⁰</i> × RK-5	<i>S⁶⁰/S⁶⁰, SRK²⁸/–/–</i>	<i>S²⁸S⁶⁰</i>	<i>S⁶⁰</i>	6	375	1.9
	<i>S⁶⁰/S⁶⁰, –/–</i>	<i>S⁶⁰</i>	<i>S⁶⁰</i>	5	104	15.3
	<i>S⁶⁰/S⁶⁰, –/–</i>	<i>S⁶⁰</i>	<i>S⁶⁰</i>	6	54	10.9
<i>S⁶⁰</i> homozygotes						
<i>S⁵²</i> × RK-5	<i>S⁵²/S⁶⁰, SRK²⁸/–/–</i>	<i>S²⁸S⁵²S⁶⁰</i>	<i>S⁵²</i>	6	312	1.1
	<i>S⁵²/S⁶⁰, –/–</i>	<i>S⁵²S⁶⁰</i>	<i>S⁵²</i>	5	118	14.9
	<i>S⁵²/S⁶⁰, –/–</i>	<i>S⁵²S⁶⁰</i>	<i>S⁵²</i>	4	42	15.9
<i>S⁵²S⁶⁰</i> heterozygotes						
LG-2-3 × RK-5	<i>S⁶⁰/S⁶⁰, SRK²⁸/–/–, SLG²⁸/–</i>	<i>S²⁸S⁶⁰</i>	<i>S⁶⁰</i>	5	383	0.3
	<i>S⁶⁰/S⁶⁰, –/–, SLG²⁸/–</i>	<i>S⁶⁰</i>	<i>S⁶⁰</i>	4	259	14.5
	<i>S²⁸/S⁶⁰, –/–</i>	<i>S²⁸S⁶⁰</i>	<i>S²⁸</i>	4	194	0.2
<i>S²⁸S⁶⁰</i> heterozygotes						

No. of plants, number of plants examined; no. of flowers, number of flowers pollinated with *S²⁸* pollen; seeds/flower, average number of seeds per flower. Three crossed progeny, *S⁶⁰* × RK-5, *S⁵²* × RK-5 and LG-2-3 × RK-5, were obtained by bud-pollinating an *S⁶⁰* homozygote, an *S⁵²* homozygote and LG-2-3 (*S⁶⁰/S⁶⁰, –/–, SLG²⁸/SLG²⁸*), respectively, with pollen from RK-5 (*S⁶⁰/S⁶⁰, SRK²⁸/–/–*). *S⁶⁰* homozygotes, *S⁵²S⁶⁰* heterozygotes and *S²⁸S⁶⁰* heterozygotes were used as controls for compatible or incompatible pollinations. S genotypes were determined by PCR. SI phenotypes of all progeny plants were determined by monitoring pollen behaviour on the stigmatic surface after pollination.

they all produced the *SRK*²⁸ transcript (Fig. 2b). Together these results indicate that the *S*²⁸ haplotype specificity acquired by the three primary transgenic plants, RK-5, RK-11 and RK-14, may have arisen from the expression of the *SRK*²⁸ transgene.

In the case of pBINSLG28 transformation, the *SLG*²⁸ transgene was detected in four independent transgenic plants, LG-1, LG-2, LG-3 and LG-4, by DNA gel blot analysis (Fig. 1c). All of them were self-incompatible (that is, rejecting *S*⁵² and *S*⁶⁰ pollen). Isoelectric focusing (IEF)-immunoblot analysis was carried out using an anti-*SLG*⁴³ monoclonal antibody that we had previously shown to crossreact with 14 of 16 class I *SLGs* examined (including *SLG*²⁸ and *SLG*⁵²), but not with 4 of class II *SLGs* (such as *SLG*⁶⁰)¹³. All these four plants produced the endogenous *SLG*⁵² protein at a normal level, but only LG-2 produced a high level of *SLG*²⁸ from the transgene (Fig. 3a). However, LG-2, as well as the other three transgenic plants, failed to reject *S*²⁸ pollen, suggesting that the stigma phenotype of these transgenic plants remained the same as that of their parental plants (*S*⁵²/*S*⁶⁰). Through bud pollination, self-pollinated progeny was raised from LG-2, and on the basis of PCR analysis, 16 out of the 18 plants carried the transgene. This segregation ratio can be explained by the mendelian inheritance of a single copy of the *SLG*²⁸ transgene ($\chi^2 = 1.7$, $P > 0.1$, d.f. = 1) carried by LG-2 (see the DNA gel blot shown in Fig. 1c). As was the case for LG-2, all 16 progeny plants that carried the *SLG*²⁸ transgene failed to reject *S*²⁸ pollen, even though they all produced high levels of *SLG*²⁸.

Among these 16 plants, LG-2-3 produced the highest level of *SLG*²⁸ protein and it was homozygous for the *SLG*²⁸ transgene (Fig. 3b). The genotype of LG-2-3 was thus designated *S*⁶⁰/*S*⁶⁰, *-/-*, *SLG*²⁸/*SLG*²⁸, with *-/-* indicating that this plant did not carry the *SRK*²⁸ transgene. LG-2-3 was crossed as female with RK-5 (*S*⁶⁰/*S*⁶⁰, *SRK*²⁸/*-*, *-/-*, with *-/-* indicating the absence of the *SLG*²⁸ transgene) by bud pollination (as both plants were homozygous for the *S*⁶⁰ haplotype and would be incompatible if crossed at the mature flower stage). The self-incompatibility phenotypes of this progeny were compared with those of the progeny from two other crosses, *S*⁶⁰ × RK-5 and *S*⁵² × RK-5. All these three progeny contained plants hemizygous for the *SRK*²⁸ transgene, because, as stated earlier, RK-5 carried one copy of the *SRK*²⁸ transgene. Without exception, the self-incompatibility phenotypes of the plants in each progeny, as determined by pollen tube behaviour, were precisely those predicted on the basis of their genotypes determined by PCR analysis (Table 1). For example, when the plants were pollinated with *S*²⁸ pollen, those that did not inherit the *SRK*²⁸ transgene yielded an average of more than 10 seeds per flower, a number comparable to that obtained from compatible crosses. In contrast, plants that inherited the *SRK*²⁸ transgene yielded an average of less than two seeds per flower.

We compared the ability of the plants in each of these three progeny to reject *S*²⁸ pollen to assess any effect that the differences in the genetic background of these three families might have on rejection of *S*²⁸ pollen. As shown in Table 1, plants of the *S*⁶⁰ × RK-5 progeny that carried the *SRK*²⁸ transgene set a few seeds (an average of 1.9 seeds per flower). Plants of the LG-2-3 × RK-5 progeny that expressed both the *SRK*²⁸ and the *SLG*²⁸ transgene (Fig. 3c) showed very strong incompatibility with *S*²⁸ pollen: an average of 0.2 seeds per flower, a number comparable to that obtained from incompatible pollination of *S*²⁸/*S*⁶⁰ heterozygotes with *S*²⁸ pollen. Plants of the *S*⁵² × RK-5 progeny that carried the *SRK*²⁸ transgene and the endogenous *SLG*⁵² and *SLG*⁶⁰ showed an intermediate level of incompatibility with *S*²⁸ pollen: an average of 1.1 seeds per flower.

Thus, the degree of *S*²⁸ pollen rejection by the plants carrying the *SRK*²⁸ transgene appears to be enhanced by the presence of the *SLG*²⁸ transgene, and to a lesser extent by the presence of the endogenous *SLG*⁵² gene, when compared with the plants carrying the *SRK*²⁸ transgene and the endogenous *SLG*⁶⁰ gene. Notably, both *S*²⁸ and *S*⁵²

belong to class I *S* haplotypes and *S*⁶⁰ belongs to class II *S* haplotypes, and the degree of *S*²⁸ pollen rejection by *SRK*²⁸ appears to correlate with the degree of amino-acid identity between its *S* domain and the *SLGs* (98% identity with *SLG*²⁸, 76% identity with *SLG*⁵² and 65% identity with *SLG*⁶⁰): the higher the identity, the stronger the rejection. As *SLGs* are secretory proteins and are abundantly present in the cell wall of the stigma²⁰, it is possible that an *SLG* has a role in the binding of its cognate *SRK* with the pollen ligand by forming a complex with the *S* domain of the *SRK* and facilitating the process of the recognition reactions. The ability of the complex formation between *SLGs* and *SRKs* might then decrease as the sequence identity between them decreases.

In conclusion, our study provides the first direct evidence to our knowledge that the *S* haplotype specificity of the stigma in self-incompatibility recognition reactions of *Brassica* is solely determined by *SRK* and that *SLG*, although not involved in *S* haplotype specificity, can enhance the process of self-incompatibility recognition reactions. These findings, coupled with the identification of the potential pollen ligand, *SCR*, in self-incompatibility interactions¹² and the finding that *ARC1*, a protein that interacts with the kinase domain of *SRK*, is required for stigma function in self-incompatibility interactions²¹, will open up opportunities for the molecular/biochemical characterization of *SRK/SCR* interactions and the *SRK*-mediated signalling pathway, which both lead to pollen rejection. In addition, our study shows that, by judicious selection of an *SRK* transgene and an *S* genotype of the recipient, it is feasible to confer a new self-incompatibility specificity on the stigma of *Brassica* plants. □

Methods

Construction of transformation vectors

*SRK*²⁸ (same as *SRK*²⁸) complementary DNA (ref. 18) and *SLG*²⁸ (same as *SLG*²⁸) genomic DNA (ref. 22) were isolated from *B. rapa* plants homozygous for the *S*²⁸ haplotype (a class I haplotype). A chimaeric gene, comprising the promoter region (3.2 kilobases (kb)) of *SLG*²⁸, the coding region of *SRK*²⁸ cDNA and the nopaline synthase transcription terminator, was inserted into a binary vector pSLJ491 (ref. 23) to yield pSLJSRK28. A 7.3-kb *HindIII* fragment of *SLG*²⁸ genomic DNA containing the promoter region (3.2 kb), the coding region, and the downstream flanking sequence (2.8 kb), was inserted into a binary vector pBin19 to generate pBINSLG28 (ref. 19).

Plant transformation

*S*⁵²/*S*⁶⁰ heterozygotes¹³ used in the transformation with pBINSLG28 were a commercial hybrid variety cv. Osome (Takii Seed Co.) of self-incompatible *B. rapa*. *S*⁶⁰ homozygous plants used in the transformation with pSLJSRK28 were obtained from bud self-pollination of Osome. We previously determined that *S*⁵² belongs to class I *S* haplotypes and *S*⁶⁰ to class II *S* haplotypes. We also established the dominant/recessive relationships among *S*²⁸, *S*⁵² and *S*⁶⁰ as follows. *S*⁵² was dominant over *S*⁶⁰ in pollen; *S*⁵² and *S*⁶⁰ were co-dominant in stigma¹³; *S*²⁸ was co-dominant with *S*⁵² and dominant over *S*⁶⁰ in pollen; *S*²⁸ was co-dominant with *S*⁵² and *S*⁶⁰ in stigma. The hypocotyl explants were transformed with *Agrobacterium tumefaciens* strain pCIB542/A136 harbouring pSLJSRK28 or pBINSLG28 according to published methods²⁴.

DNA gel blot analysis

Total DNA was extracted from young leaves. Two µg DNA digested with *HindIII*, *BamHI* or *XbaI* were electrophoresed on 0.8% agarose gels and transferred to nylon membranes for hybridization with digoxigenin (Dig)-labelled cDNA encoding the kinase domain of *SRK*²⁸ (*SRK*²⁸-KD) or *SLG*²⁸ (ref. 18), according to the manual of the Dig Nucleic Acid Detection Kit (Boehringer Mannheim). After hybridization, the membranes were washed twice in 0.1% SSC, 0.1% SDS at 65 °C for 20 min.

Pollination test

The pollination data are based on pollen tube counts that were determined by ultraviolet-fluorescence microscopy²⁵ with more than 30 flowers in respective pollinations.

RNA gel blot analysis

For RNA gel blot analysis, poly(A)⁺ RNA was isolated from stigmas and anthers at one day before anthesis by using the Micro Fast Track messenger RNA Isolation Kit (Invitrogen). After denaturation in glyoxal, the RNA was loaded on 1% agarose gels in 10 mM sodium phosphate buffer pH 7.0 for electrophoresis. For the *S*²⁸ homozygote, 1 µg of poly(A)⁺ RNA was loaded; for all the other plants, 2 µg of poly(A)⁺ RNA was loaded. The RNA was transferred to nylon membranes and hybridized with Dig-labelled *SRK*²⁸-KD, *SLG*²⁸ cDNA, *SLG*⁶⁰ PCR fragment¹³ and vacuolar H⁺ ATPase (*V-ATPase*) fragment. The

V-ATPase fragment was amplified from the genomic DNA of Osome plants by PCR using the primers of its conserved region²⁶. Washing and detection were carried out as described for DNA gel blot analysis.

Detection of transgene by PCR

S-genotypes of self-pollinated and crossed progeny were determined as follows. Genomic DNA was prepared from young leaves. The *SLG*²⁸ transgene was amplified by PCR with *SLG*²⁸ specific primers, PS18 and PS15 (ref. 27). The *SRK*²⁸ transgene was amplified by PCR with *SRK*²⁸ specific primers, PK28 (5'-CCTCTTATTTTCTGCTCTGG-3') and PK4 (ref. 28). PK28 was designed on the basis of the nucleotide sequence of the transmembrane domain of *SRK*²⁸, and PK4 was designed on that of exon 4 of *SRK*²⁸. The expected PCR product of the *SRK*²⁸ gene was 648 bp. To discriminate between the *SRK*²⁸ transgene and the endogenous *SRK*²⁸ gene, PCR-RFLP was conducted by using the primers, PS18 and B (ref. 29) to obtain PCR products, which were then digested with *Mbo*I and electrophoresed on 5% polyacrylamide gels; the fragments were visualized by silver-staining. The endogenous *SLG*⁵² and *SLG*⁶⁰ genes were amplified by PCR with PS5 and PS15, and PS3 and PS21, respectively^{13,27}.

Immunoblot analysis

Total protein was extracted from five stigmas for the *S*²⁸ homozygote and ten stigmas for all the other plants in 50 mM Tris-HCl pH 7.5. The extract was subjected to thin-layer polyacrylamide gel IEF (Ampholine PAG Plate, pI 3.5–9.5; LKB Pharmacia) and transferred to PVDF membranes (Millipore) by electroblotting. SLG proteins were detected with the anti-SLG⁴³ (a class I SLG) monoclonal antibody which crossreacts with most of class I SLGs (ref. 30).

Received 21 October; accepted 14 December 1999.

- Bateman, A. J. Self-incompatibility systems in angiosperms. III. Cruciferae. *Heredity* **9**, 52–68 (1955).
- Stein, J. C. et al. Molecular cloning of a putative receptor protein kinase gene encoded at the self-incompatibility locus of *Brassica oleracea*. *Proc. Natl Acad. Sci. USA* **88**, 8816–8820 (1991).
- Takayama, S. et al. Sequences of S-glycoproteins, products of the *Brassica campestris* self-incompatibility locus. *Nature* **326**, 102–105 (1987).
- Nasrallah, J. B., Kao, T.-h., Chen, C.-h., Goldberg, M. L. & Nasrallah, M. E. Amino-acid sequence of glycoproteins encoded by three alleles of the S-locus of *Brassica oleracea*. *Nature* **326**, 617–619 (1987).
- Goring, D. R., Glavin, T. L., Schafer, U. & Rothstein, S. J. An S receptor kinase gene in self-compatible *Brassica napus* has a 1-bp deletion. *Plant Cell* **5**, 531–539 (1993).
- Nasrallah, J. B., Rundle, S. J. & Nasrallah, M. E. Genetic evidence for the requirement of the *Brassica* S-locus receptor kinase gene in the self-incompatibility response. *Plant J.* **5**, 373–384 (1994).
- Nasrallah, M. E., Kandasamy, M. K. & Nasrallah, J. B. A genetically defined *trans*-acting locus regulates S-locus function in *Brassica*. *Plant J.* **2**, 497–506 (1992).
- Goring, D. R. & Rothstein, S. J. The S-locus receptor kinase gene in a self-incompatible *Brassica napus* line encodes a functional serine/threonine kinase. *Plant Cell* **4**, 1273–1281 (1992).
- Nasrallah, J. B. & Nasrallah, M. E. Pollen-stigma signaling in the sporophytic self-incompatibility response. *Plant Cell* **5**, 1325–1335 (1993).
- Hinata, K., Watanabe, M., Toriyama, K. & Isogai, A. A review of recent studies on homomorphic self-incompatibility. *Int. Rev. Cytol.* **143**, 257–296 (1993).
- Suzuki, G. et al. Genomic organization of the S locus: Identification and characterization of genes in *SLG*/*SRK* region of *S*²⁸ haplotype of *Brassica campestris* (syn. *rapa*). *Genetics* **153**, 391–400 (1999).
- Schopfer, C. R., Nasrallah, M. E. & Nasrallah, J. B. The male determinant of self-incompatibility in *Brassica*. *Science* **286**, 1697–1700 (1999).
- Takasaki, T. et al. Introduction of *SLG* (S-locus glycoprotein) alters the phenotype of endogenous S-haplotype, but confers no new S-haplotype specificity in *Brassica rapa* L. *Plant Mol. Biol.* **40**, 659–668 (1999).
- Conner, J. A. et al. Transgene-induced silencing of S-locus genes and related genes in *Brassica*. *Plant J.* **11**, 809–823 (1997).
- Stahl, R. J., Arnold, M., Glavin, T. L., Goring, D. R. & Rothstein, S. J. The self-incompatibility phenotype is altered by the transformation of a mutant S locus receptor kinase. *Plant Cell* **10**, 209–218 (1998).
- Hatakeyama, K., Takasaki, T., Watanabe, M. & Hinata, K. Molecular characterization of S locus genes, *SLG* and *SRK*, in a pollen-recessive self-incompatibility haplotype of *Brassica rapa* L. *Genetics* **149**, 1587–1597 (1998).
- Hatakeyama, K., Watanabe, M., Takasaki, T., Ojima, K. & Hinata, K. Dominance relationships between S alleles in self-incompatible *Brassica campestris* L. *Heredity* **80**, 241–247 (1998).
- Watanabe, M. et al. A high degree of homology exists between the protein encoded by *SLG* and the S receptor domain encoded by *SRK* in self-incompatible *Brassica campestris* L. *Plant Cell Physiol.* **35**, 1221–1229 (1994).
- Suzuki, G., Watanabe, M., Toriyama, K., Isogai, A. & Hinata, K. Expression of *SLG*²⁸ and *SRK*²⁸ genes in transgenic tobacco. *Plant Cell Physiol.* **37**, 866–869 (1996).
- Kish-Nishizawa, N. et al. Ultrastructure of papillar cells in *Brassica campestris* revealed by liquid helium rapid-freezing and substitute-fixation method. *Plant Cell Physiol.* **31**, 1207–1219 (1990).
- Stone, S. L., Arnold, M. & Goring, D. R. A breakdown of *Brassica* self-incompatibility in ARC1 antisense transgenic plants. *Science* **286**, 1729–1731 (1999).
- Suzuki, G., Watanabe, M., Toriyama, K., Isogai, A. & Hinata, K. Molecular cloning of members of the S-multigene family in self-incompatible *Brassica campestris*. *Plant Cell Physiol.* **36**, 1273–1280 (1995).
- Jones, J. G. et al. Effective vectors for transformation, expression of heterologous, and assaying transposon excision in transgenic plant. *Transgenic Res.* **1**, 285–297 (1992).
- Takasaki, T. et al. Factors influencing *Agrobacterium*-mediated transformation of *Brassica rapa* L. *Breed. Sci.* **47**, 127–134 (1997).
- Kho, Y. O. & Bear, J. Observing pollen tubes by means of fluorescence. *Euphytica* **17**, 298–302 (1968).
- Manolson, M. F., Ouellette, B. F., Filion, M. & Poole, R. J. cDNA sequence and homologies of the '57-Kda' nucleotide-binding subunit of the vacuolar ATPase from *Arabidopsis*. *J. Biol. Chem.* **263**, 17987–17994 (1988).

- Nishio, T., Kusaba, M., Watanabe, M. & Hinata, K. Registration of S alleles in *Brassica campestris* L. by the restriction fragment sizes of *SLGs*. *Theor. Appl. Genet.* **92**, 388–394 (1996).
- Nishio, T., Kusaba, M., Sakamoto, K. & Ockendon, D. Polymorphism of the kinase domain of the S-locus receptor kinase gene (*SRK*) in *Brassica oleracea* L. *Theor. Appl. Genet.* **95**, 335–342 (1997).
- Brace, J., Ockendon, D. J. & King, G. J. Development of a method for identification of S alleles in *Brassica oleracea* based on digestion of PCR-amplified DNA with restriction endonucleases. *Sex. Plant Reprod.* **6**, 133–138 (1993).
- Shiba, H., Hinata, K., Suzuki, A. & Isogai, A. Breakdown of self-incompatibility in *Brassica* by the antisense RNA of the *SLG* gene. *Proc. Jpn. Acad. B* **71**, 81–83 (1995).

Acknowledgements

We thank T.-h. Kao for his critical review and editing of the manuscript. This work was supported in part by Grants-in-Aid for Special research on Priority Areas from the Ministry of Education, Science, Culture and Sports, Japan.

Correspondence and requests for materials should be addressed to T.T. (e-mail: taka@kobe-u.ac.jp).

Structure of the winged-helix protein hRFX1 reveals a new mode of DNA binding

Ketan S. Gajiwala*, Hua Chen*†, Fabrice Cornille‡, Bernard P. Roques‡, Walter Reith§, Bernard Mach§ & Stephen K. Burley*†

*Laboratories of Molecular Biophysics, Pels Family Center for Biochemistry and Structural Biology, †Howard Hughes Medical Institute, The Rockefeller University, 1230 York Avenue, New York, New York 10021, USA
‡Département de Pharmacologie Moléculaire et Structurale, INSERM U266, CNRS URA D1500, UFR des Sciences Pharmaceutiques et Biologiques, 4 Avenue de l'Observatoire, 75720 Paris Cedex 06, France
§Département de Génétique et Microbiologie, Centre Médical Universitaire (CMU), 1, rue Michel-Servet, CH 1211 Genève, 4 Switzerland

Regulatory factor X (RFX) proteins are transcriptional activators that recognize X-boxes (DNA of the sequence 5'-GTNRCC(0–3N)RGYAAAC-3', where N is any nucleotide, R is a purine and Y is a pyrimidine) using a highly conserved 76-residue DNA-binding domain (DBD). DNA-binding defects in the protein RFX5 cause bare lymphocyte syndrome or major histocompatibility antigen class II deficiency¹. RFX1, -2 and -3 regulate expression of other medically important gene products (for example, interleukin-5 receptor α chain, IL-5R α)². Fusions of the ligand-binding domain of the oestrogen receptor with the DBD of RFX4 occur in some human breast tumours³. Here we present a 1.5 Å-resolution structure of two copies of the DBD of human RFX1 (hRFX1) binding cooperatively to a symmetrical X-box^{4,5}. hRFX1 is an unusual member of the winged-helix subfamily of helix–turn–helix proteins⁶ because it uses a β -hairpin (or wing) to recognize DNA instead of the recognition helix typical of helix–turn–helix proteins. A new model for interactions between linker histones and DNA is proposed.

We used multiwavelength anomalous dispersion (MAD) to determine the structure of the hRFX1 DBD (Fig. 1) recognizing a symmetrical X-box (see Methods). Unexpectedly, the hRFX1 DBD proved to be a member of the winged-helix hepatocyte nuclear factor (HNF)-3/forkhead-related subfamily of HTH transcription factors (Fig. 2, reviewed in ref. 7). The DBD consists of three α -helices (H), three β -strands (S) and three connecting loops (L), arranged in the order H1-S1-H2-L1-H3-L2-S2-W1-S3. The third loop, connecting β -strands S2 and S3, forms wing W1 of the winged-helix motif. Residues forming the hydrophobic core of the DBD originate from the secondary structural elements H1, S1, H2,

H3, L2 and S3 (Fig. 1). Fifteen out of nineteen of these amino acids are either identical or conserved among RFX proteins, indicating that all five distinct DBDs are likely to share the same structure. In contrast to previously described winged-helix DBDs (typically 110

residues long with two wings, W1 and W2)⁶, the shorter RFX DBD has only one wing (Fig. 2). A structure-based sequence alignment of the hRFX1 DBD with HNF-3 γ revealed 8% identity (Fig. 2b), which explains why we could not discern their similarity. As deduced from

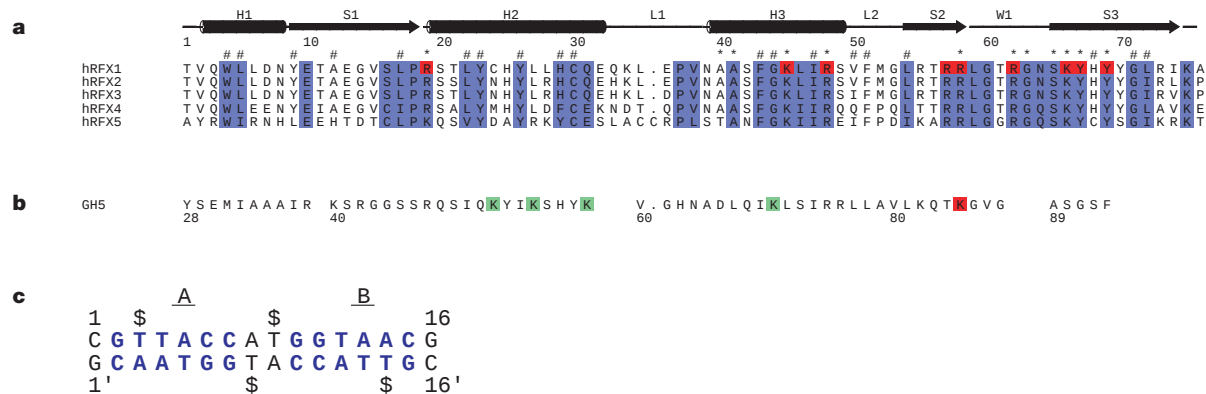


Figure 1 Protein and oligonucleotide sequences. **a**, DBDs of five RFX proteins with the hRFX1 DBD secondary structure; pairwise identities range between 36 and 90%. Violet denotes conserved amino acids; red identifies hRFX1 residues making direct or water-mediated DNA contacts. Residues marked with hash signs comprise the hydrophobic core, and side chains buried on DNA binding are indicated with asterisks. Residue numbering is based on the hRFX1 DBD with Thr 1 corresponding to Thr 438 in full-length

hRFX1. **b**, Structure-based sequence alignment of GH5 with the hRFX1 DBD. Green coded GH5 residues are protected against chemical modification in chromatin, with Lys 85 (red) being maximally protected. **c**, Oligonucleotide used for crystallization with numbering scheme, X-box half-site labelling and 5-bromouracil positions (dollar signs). Purple lettering denotes half-sites.

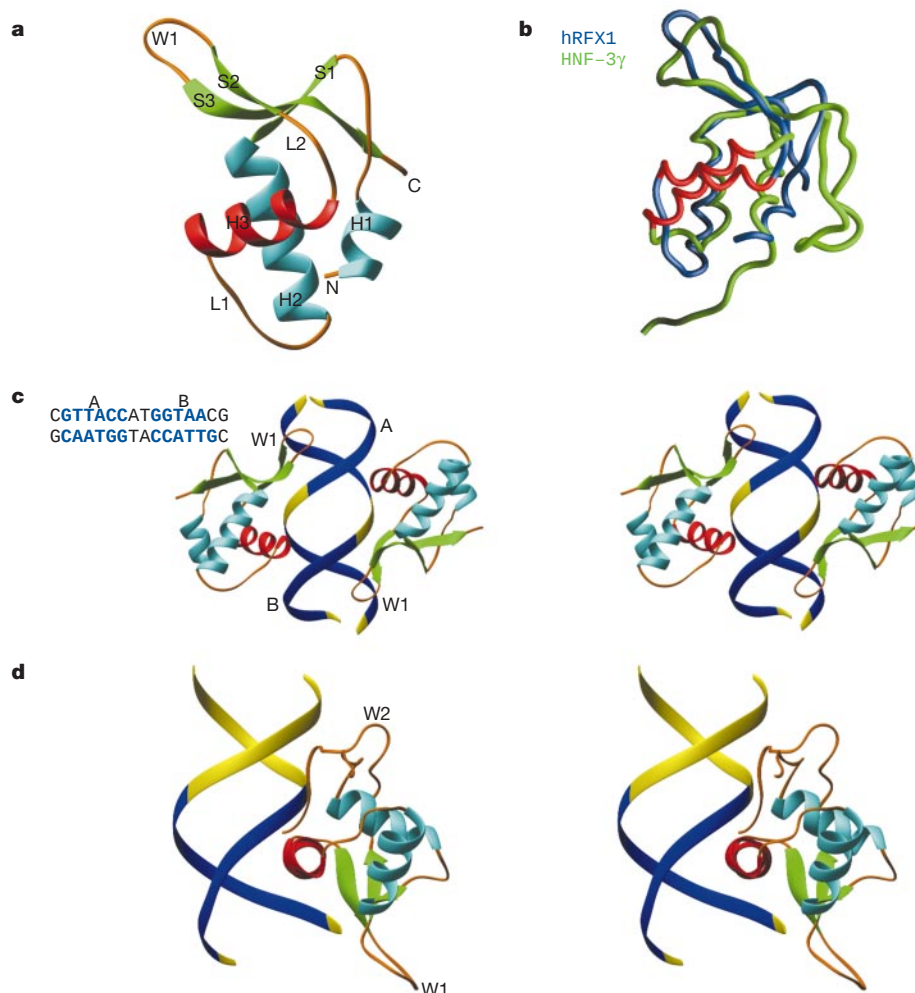


Figure 2 hRFX1 and HNF-3 γ DBDs. **a**, hRFX1 DBD with labelled N and C termini and secondary structural elements (HTH recognition helix H3, red). **b**, Superposition of hRFX1 and HNF-3 γ DBDs (r.m.s. deviation for α -carbon atoms, 3.5 Å), shown in the same orientation as **a**. **c**, Stereo diagram of the hRFX1–DNA 2:1 complex, viewed along the

2-fold crystallographic symmetry axis relating each half of the asymmetric unit. Inset, crystallization oligonucleotide, with blue characters denoting X-box half-sites, labelled A and B. **d**, Stereo diagram of HNF-3 γ bound to its cognate DNA⁶.

previous biochemical studies⁵, two RFX1 DBDs bind a single X-box. The resulting 2:1 protein–DNA complex is perfectly symmetrical, with a crystallographic two-fold axis passing through the centre of the DNA (Fig. 2c).

The hRFX1 consensus binding site is an imperfect inverted repeat with variable spacing between two half-sites (5'-NNGTNRCC/n(0–3N)RGYAACNN-3'; underlining denotes X-box half-site)⁸. The palindromic 16-base pair (bp) duplex oligonucleotide used here (5'-CGTTACCATGGTAACG-3') has a 2-bp spacer (5'-AT-3'). This X-box closely resembles the physiological hRFX1 binding site (5'-GTTGCCCGGCAAC-3') in the hepatitis B virus (HBV) enhancer⁹. Each half-site of the symmetric X-box interacts with both DBDs. Protein–DNA interactions are shown in Figs 2 and 3. Among previously reported structures of *bona fide* winged-helix proteins (that is, those sharing the HNF-3 γ fold), four have been determined with cognate B-DNAs (HNF-3 γ ⁶, genesis¹⁰, E2F4 and DP2 (ref. 11)). In the HNF-3 γ co-crystal structure, H3 (that is, the recognition helix; Fig. 2d, red) makes specific contacts with the major groove⁶. W1 of HNF-3 γ overlies the minor groove of a neighbouring oligonucleotide in the crystal lattice, and the carboxy-terminal wing (W2) makes one minor-groove contact (Fig. 2d). This mode of DNA binding is shared by the transcription factors genesis, E2F4 and DP2, and is common to all fully characterized HTH proteins except hRFX1.

The hRFX1 DBD uses an alternative mechanism of DNA recognition. β -strands S2 and S3 and their connecting loop, wing W1, make extensive contacts with the major groove of one half-site of the

symmetric X-box (Figs 2c and 3). A single side chain from α -helix H3 (red in Figs 2c and 3) interacts with the minor groove of the other half-site. X-box binding by hRFX1 is mediated primarily by polar side chains, through direct contacts or via interfacial water molecules (Fig. 3). This unprecedented protein–DNA complex is not a crystal packing artefact. The total solvent-accessible surface area buried per DBD is 1,800 Å², which far exceeds values characteristic of adventitious lattice contacts and is compatible with a specific protein–DNA complex (reviewed in ref. 12). The observed side-chain–base contacts (Fig. 3) are entirely consistent with hRFX1 binding *in vivo* and *in vitro*.

Binding of hRFX1 to the Py site in the polyoma virus genome is abolished by changing G11 to any other base⁸. This finding can be explained by our structure (Fig. 3a), which contains two hydrogen bonds between Arg 58 and G11 (NH1–O6 = 2.7 Å, NH2–N7 = 3.1 Å). Replacement of G11 with A would eliminate the NH1–O6 hydrogen bond, and both hydrogen bonds would probably be impossible following pyrimidine substitution. Our explanation is supported by the results of electrophoretic gel mobility shift assays (EMSA). When G11 and G10 are changed to T simultaneously, eliminating three hydrogen bonds (G10 interacts with Tyr 67, OH–N7 = 2.7 Å, not depicted), DNA binding is abolished (data not shown). A similar effect is observed when G15' is substituted with T, presumably because hydrogen bonds with Arg 62 (Fig. 3b, NE–N7 = 3.1 Å, NH2–O6 = 2.7 Å) can no longer form (data not shown). Additional, albeit indirect, support for the physiological relevance of our co-crystal structure comes from studies of RFX proteins

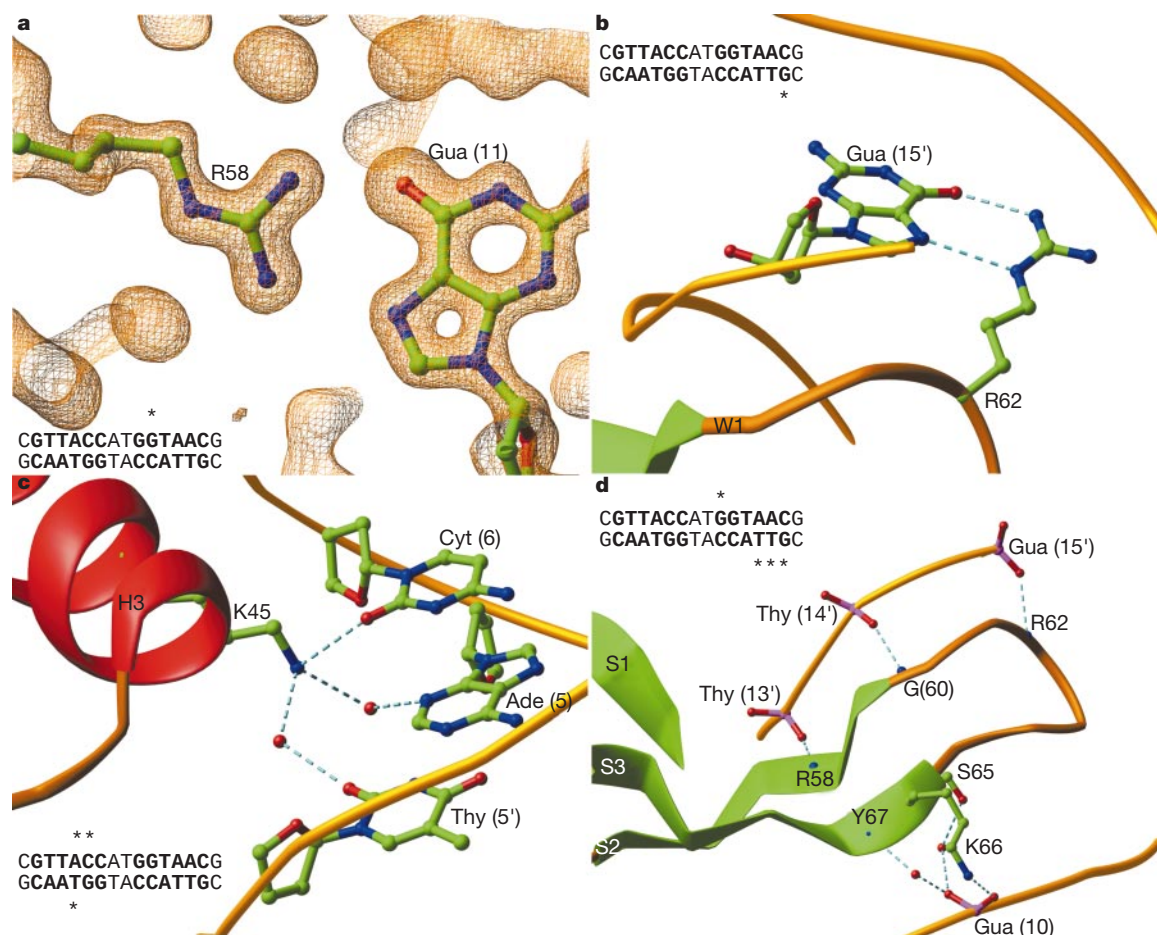


Figure 3 Protein–DNA interactions. **a**, Representative 1.5 Å resolution electron-density map contoured at 2 σ , showing Arg 82 bound to G11 in the major groove of the B half-site. The crystallization oligonucleotide is inset, with bold denoting each half-site, and an asterisk identifying the depicted nucleotide(s). **b**, Arg 62 bound to G15' in the major

groove of the B half-site. **c**, Lys 45 from H3 (red) interacts with the minor groove face of the A half-site. Red spheres denote water molecules. **d**, Direct and water-mediated protein–DNA contacts made by W1 with the phospho-deoxyribose backbones, flanking the major groove of the B half-site.

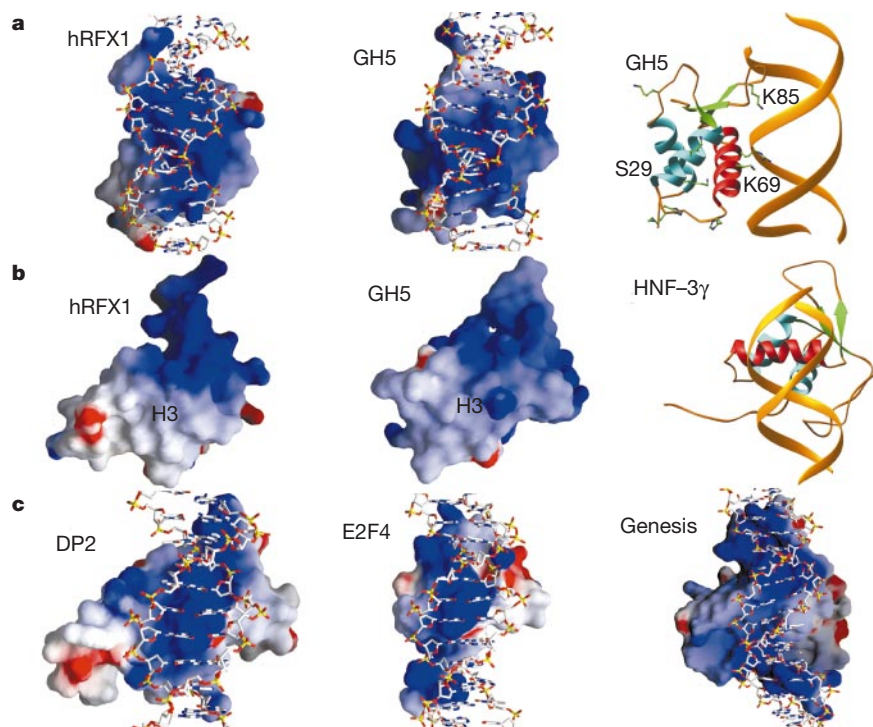


Figure 4 Model for linker histone–DNA interactions. GRASP²⁵ representations of the chemical properties of the solvent-accessible DBD surface, calculated using a water probe radius of 1.4 Å. The surface electrostatic potential is coloured red and blue, representing electrostatic potentials <-20 to $>+20k_B T$, where k_B is the Boltzmann constant and T is the temperature. The calculations were performed with an ionic strength of 0 and dielectric constants of 80 and 2 for solvent and protein, respectively. DNA is denoted as an atomic stick figure or a pair of yellow ribbons denoting phospho-deoxyribose backbones. **a**, hRFX1 DBD–DNA complex (left), and our alternative model of

GH5–DNA complex (middle, surface electrostatic potential; right, ribbon diagram). Depicted GH5 side chains (right) are thought to be important for DNA packaging. Label denoting Ser 29 shows its approximate, partially obscured position. **b**, H3 faces colour coded for electrostatic potential (left: hRFX1 DBD; middle: GH5). The HNF-3 γ –DNA ribbon drawing (right) shows the horizontal orientation of H3 common to all views in **b** and **c**. **c**, DP2, E2F4 and genesis bound to their cognate DNAs, with H3 faces colour-coded for electrostatic potential.

binding to X-boxes containing methylated 5′-CpG-3′ dinucleotides (consensus sequence for RFX binding to methylated sites is 5′-N RTMRY^YAMRGMRAYN-3′, where R is a purine, Y is a pyrimidine and M is 5-methyl-Cyt or T¹³). Methyl groups of T4 and T6, respectively, make van der Waals contacts with Arg 58 (Me–CB = 4 Å, not shown) and Tyr 67 (Me–CZ = 3.6 Å, not shown).

Finally, studies of the IL-5R α promoter show that our findings with RFX1 can be generalized to other members of the RFX family. Expression of IL-5R α is regulated by RFX1, -2 and -3 through homomeric and heteromeric interactions with a unique enhancer-like *cis* element, which resembles the consensus RFX binding site (5′-NNGTNRCC/n(0-3N)RGYAACNN-3′)². Mutations of three guanines within the IL-5R α promoter completely abolish binding of RFX1, -2, and -3 (ref. 2). These deleterious substitutions involve

nucleotides making major groove hydrogen bonds with Arg 58, Arg 62 and Tyr 67 in our RFX1 co-crystal structure. It is, therefore, likely that RFX2 and -3 bind to the IL-5R α X-box using the same mechanism as RFX1, because all nine of the residues involved in DNA contacts are identical among these RFX family members.

Our structure of the hRFX1–DNA complex reveals no protein–protein interactions between the two DBDs comprising the 2:1 protein–DNA complex (Fig. 2c), which readily explains how RFX1, -2 and -3 can form heterocomplexes with the IL-5R α X-box. EMSA has revealed cooperative hRFX1 DBD binding to an X-box⁵. These observations indicate that DNA-binding cooperativity may arise from a through-DNA effect. Geometric analysis of X-box DNA conformational parameters¹⁴ reveals that the geometry of each X-box half-site is distorted when bound to the hRFX1 DBD. The

Table 1 Statistics of the Crystallographic Analysis

Data set	Resolution (Å)	Reflections, measured/unique	Completeness (%), overall/outer shell	R_{sym} (%), overall/outer shell
λ_1 (0.9201 Å)	20.0–1.85	112,611/12,998	99.3/93.7	4.0/16.4
λ_2 (0.9197 Å)	20.0–1.85	112,848/13,014	99.3/93.7	4.2/18.0
λ_3 (0.9062 Å)	20.0–1.85	114,270/12,998	99.8/98.3	4.2/18.8
Overall MAD figure of merit, 0.61				
Refinement statistics	Resolution (Å)	R_{sym} /completeness (%)	R factor, overall/outer shell	Free R factor
Data	10.0–1.5	5.2/96.3	0.192/0.216	0.227
r.m.s. deviations	bond lengths, 0.016 Å	bond angles, 1.8°	thermal parameters, 1.3 Å ²	

$R_{\text{sym}} = \sum_i |I_i - \langle I \rangle| / \sum_i I_i$, where I_i is observed intensity and $\langle I \rangle$ is average intensity obtained from multiple observations of symmetry related reflections.

r.m.s. bond lengths and r.m.s. bond angles are the respective root-mean-square deviations from ideal values. r.m.s. Thermal parameter is the root-mean-square deviation between the B values of covalently bonded atomic pairs. Free R factor was calculated with 10% of the data omitted from the structure refinement.

minor groove is widened (by $>3 \text{ \AA}$) where H3 inserts Lys 45 to make a hydrogen bond with C6 (Fig. 3c). On the opposite face of the double helix, where the other DBD presents W1 to the major groove, narrowing is observed (by $1 \text{ \AA}$). Hence major groove recognition of an X-box half-site by one DBD yields a narrowed major groove conformation, which in turn could stabilize binding of the second DBD because minor groove widening would facilitate insertion of Lys 45 (or vice versa).

During cooperative binding of two RFX DBDs to an X-box, reciprocal interactions would occur between both DBDs and the A and B half-sites (Figs 1c and 2c). An alternative explanation of DNA-binding cooperativity is provided by the presence of overlapping protein–DNA contacts. In the A half-site, Ser 42 of one DBD makes a water-mediated, major groove contact with G6' (not shown), whereas Lys 45 of the other DBD in the 2:1 complex makes a direct minor groove hydrogen bond with C6 on the opposite strand (Fig. 3c). Corresponding contacts are seen in the B half-site with G11 and C11'. A comparable overlap has been used to explain cooperativity in the Oct1 POU domain–DNA complex¹⁵. It is not possible to distinguish the relative contributions of these two effects in the absence of an exhaustive analysis of hRFX1 DNA-binding cooperativity.

An X-ray structure of the globular portion of histone H5 (GH5) was determined in the absence of DNA¹⁶. Recognizing that GH5 was a HTH protein, biochemical data^{17,18} allowed the authors¹⁶ to suggest that H3 of GH5 functions as a recognition helix, which would interact with the major groove of DNA in chromatin. Our hRFX1 co-crystal structure reveals a second mechanism by which the linker histones H1 and H5 may bind DNA. Using a structure-based sequence alignment¹⁹, GH5 was superimposed on the hRFX1 DBD–DNA complex (Fig. 4a, left), yielding the structural model for linker histones binding to DNA depicted in Fig. 4a (middle and right). This alternative W1/major-groove binding model is consistent with available biochemical data. Lys 85 in GH5 (Arg 58 in hRFX1 DBD, Fig. 1b) is essential for DNA binding and is protected from reductive methylation in chromatin¹⁷. Mutation of Lys 85 to either Glu or Gln renders GH5 unable to protect nucleosomal DNA from enzymatic digestion¹⁸. In the model shown in Fig. 4a (right), GH5 of Lys 85 can hydrogen bond to G11, which represents an important contact site for the hRFX1 DBD (Fig. 3a). Another basic amino acid in GH5 (Lys 69, Fig. 4a) has been implicated in DNA binding¹⁷. The corresponding residue in the hRFX1 DBD (Lys 45, Fig. 3c) extends from H3 into the minor groove of the X-box and interacts with C6. Finally, photoaffinity-labelling studies place His 62 of GH5 close to DNA²⁰, which is also compatible with the model shown in Fig. 4a. These indirect biochemical observations do not allow us to choose between the H3/major groove and W1/major groove models for linker histone–DNA interactions, because H3 and W1 interact with DNA in both the HNF-3 γ and the hRFX1 co-crystal structures.

The electrostatic properties of the winged-helix proteins provide a means of resolving the issue. Examination of solvent-accessible surfaces of the hRFX1 DBD and GH5 reveals that the W1 face of GH5 (that is, the DNA-binding surface of hRFX1) is significantly more basic than its H3 face (compare Fig. 4a and 4b, middle). The same is true for the hRFX1 DBD (compare Fig. 4a and 4b, left). (The basic patch immediately adjacent to the hRFX1 label in Fig. 4b is located in the background and does not sit in the same plane as the hydrophobic face of H3.) In contrast, the H3 faces of HNF-3 γ , DP2, E2F4 and genesis (Fig. 4c) are significantly more basic than their respective W1 faces (not depicted). Winged-helix proteins can, therefore, be divided into two functional classes on the basis of their electrostatic properties. HNF-3 γ , DP2, E2F4 and genesis bind DNA with basic H3 faces, placing their recognition helices in the major groove. hRFX1 uses a basic W1 face for DNA binding. We believe that the striking similarity of the electrostatic properties of the hRFX1 DBD and GH5 provides compelling support for the

model of DNA binding by linker histones shown in Fig. 4a (middle, right). This model is not incompatible with the fact that a murine HNF-3 can functionally substitute for the linker histone H1 in the mouse prealbumin gene enhancer²¹, because we do not know how HNF-3 α interacts with DNA in the presence of histones H2a, H2b, H3 and H4, and other transcription factors. Moreover, the model does not preclude another portion of GH5 acting as a secondary DNA-binding site, as characterized in ref. 20. Ser 29 of GH5 (Fig. 4a, right) participates in DNA binding²², and may well represent part of this secondary nucleic acid binding surface.

In conclusion, our 1.5 Å resolution co-crystal structure of the hRFX1 DBD reveals an unprecedented mode of DNA recognition for an HTH protein. This result provides insight into DNA binding by linker histones, and demonstrates the functional versatility of the winged-helix DNA-binding motif. Finally, this structure provides a framework for further studies of the bare lymphocyte syndrome and other medically important roles of the RFX proteins. □

Methods

X-ray crystallography

Synthetic hRFX1 DBD⁵ bearing two substitutions, Cys 24 Asn and Cys 30 Ser, was crystallized with a synthetic brominated 16-bp oligonucleotide (Fig. 1c). Co-crystals (space group C2: $a = 72.8 \text{ \AA}$, $b = 42.8 \text{ \AA}$, $c = 52.3 \text{ \AA}$, $\beta = 109.2^\circ$) with one DBD plus one DNA strand per asymmetric unit, were grown by hanging drop vapour diffusion at 4°C against 20% PEG-400, 80 mM magnesium acetate, 50 mM cacodylate, pH 6.5. MAD data were collected under cryogenic conditions using CHES beamline F2, and the refinement data were collected using Brookhaven NSLS beamline X9B (Table 1).

Structure determination and refinement

Both bromines were located using anomalous difference Patterson syntheses. MAD phases at 1.85 Å resolution yielded an interpretable experimental electron-density map. After model building, the structure was refined to convergence at 1.5 Å resolution using CNS²³. The final refinement model includes hRFX1 DBD residues 1–76, the crystallization DNA, 124 water molecules, 2 molecules of ethylene glycol and 1 molecule of diethylene glycol (Table 1). PROCHECK²⁴ revealed no unfavourable (ϕ , ψ) combinations in the hRFX1 DBD, with main-chain and side-chain structural parameters consistently better than average (overall G value 0.1).

Received 6 September 1999; accepted 5 January 2000.

- Mach, B., Steimle, V., Martinez-Soria, E. & Reith, W. Regulation of MHC class II genes: lessons from a disease. *Annu. Rev. Immunol.* **14**, 301–331 (1996).
- Iwama, A. *et al.* Dimeric RFX proteins contribute to the activity and lineage specificity of the interleukin-5 receptor alpha promoter through activation and repression domains. *Mol. Cell Biol.* **19**, 3940–3950 (1999).
- Dotzlaw, H., Alkhalaf, M. & Murphy, L. C. Characterization of estrogen receptor variant mRNAs from human breast cancers. *Mol. Endocrinol.* **6**, 773–785 (1992).
- Reith, W. *et al.* MHC class II regulatory factor RFX has a novel DNA-binding domain and a functionally independent dimerization domain. *Genes Dev.* **4**, 1528–1540 (1990).
- Cornille, F. *et al.* DNA binding properties of a chemically synthesized DNA binding domain of hRFX1. *Nucleic Acids Res.* **26**, 2143–2149 (1998).
- Clark, K. L., Halay, E. D., Lai, E. & Burley, S. K. Co-crystal structure of the HNF-3/fork head DNA-recognition motif resembles histone H5. *Nature* **364**, 412–420 (1993).
- Lai, E., Clark, K. L., Burley, S. & James E. Darnell, J. Hepatocyte nuclear factor 3/fork head or "winged helix" proteins: A family of transcription factors of diverse biological function. *Proc. Natl. Acad. Sci. USA* **90**, 10421–10423 (1993).
- Emery, P. *et al.* A consensus motif in the RFX DNA binding domain and binding domain mutants with altered specificity. *Mol. Cell Biol.* **16**, 4486–4494 (1996).
- Ostapchuk, P., Scheirle, G. & Hearing, P. Binding of nuclear factor EF-C to a functional domain of the hepatitis B virus enhancer region. *Mol. Cell Biol.* **9**, 2787–2797 (1989).
- Jin, C., Marsden, L., Chen, X. & Liao, X. Dynamic DNA contacts observed in the NMR structure of winged helix protein–DNA complex. *J. Mol. Biol.* **289**, 683–690 (1999).
- Zheng, N., Fraenkel, E., Pabo, C. O. & Pavletich, N. P. Structural basis of DNA recognition by the heterodimeric cell cycle transcription factor E2F-DP. *Genes Dev.* **13**, 666–674 (1999).
- Janin, J. Principles of protein–protein recognition from structure to thermodynamics. *Biochimie* **77**, 497–505 (1995).
- Reith, W. *et al.* RFX1, a transactivator of hepatitis B virus enhancer I, belongs to a novel family of homodimeric and heterodimeric DNA-binding proteins. *Mol. Cell Biol.* **14**, 1230–1244 (1994).
- Lavery, R. & Sklenar, H. The definition of generalized helicoidal parameters and of axis curvature for irregular nucleic acids. *J. Biomol. Struct. Dynamics* **6**, 63–91 (1988).
- Klemm, J. D. & Pabo, C. O. Oct-1 POU domain–DNA interactions: cooperative binding of isolated subdomains and effects of covalent linkage. *Genes Dev.* **10**, 27–36 (1996).
- Ramakrishnan, V., Finch, J., Graziano, V. & Sweet, R. Crystal structure of the globular domain of histone H5 and its implications for nucleosome binding. *Nature* **362**, 219–223 (1993).
- Thomas, J. O. & Wilson, C. M. Selective radiolabelling and identification of a strong nucleosome binding site on the globular domain of histone H5. *EMBO J.* **5**, 3531–3537 (1986).
- Buckle, R. S., Maman, J. D. & Allan, J. Site-directed mutagenesis studies on the binding of the globular domain of linker histone H5 to the nucleosome. *J. Mol. Biol.* **223**, 651–659 (1992).

19. Holm, L. & Sander, C. Families of structurally similar proteins, version 1.0. *J. Mol. Biol.* **233**, 123–138 (1993).
20. Goytisolo, F. A. *et al.* Identification of two DNA-binding sites on the globular domain of histone H5. *EMBO J.* **15**, 3421–3429 (1996).
21. Cirillo, L. A. *et al.* Binding of the winged-helix transcription factor HNF3 to a linker histone site on the nucleosome. *EMBO J.* **17**, 244–254 (1998).
22. Zhou, Y. B., Gerchman, S. E., Ramakrishnan, V., Travers, A. & Muyldermans, S. Position and orientation of the globular domain of linker histone H5 on the nucleosome. *Nature* **395**, 402–405 (1998).
23. Brunger, A. T. *et al.* Crystallography and NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D Biol. Crystallogr.* **54**, 905–921 (1998).
24. Laskowski, R. J., MacArthur, M. W., Moss, D. S. & Thornton, J. M. PROCHECK: a program to check stereochemical quality of protein structures. *J. Appl. Cryst.* **26**, 283–290 (1993).
25. Nicholls, A., Sharp, K. & Honig, B. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins Struct. Funct. Genet.* **11**, 281–296 (1991).

Acknowledgements

We thank D. Thiel and the MacCHESS staff, and Z. Dauter and K.R. Rajashankar and the NSLS X9B staff for help with X-ray data collection. We thank C. Lenoir for assistance in peptide synthesis, and J. Bonanno, K. L. Clark, J. E. Darnell, R. C. Deo, D. Jeruzalmi, J. Kuriyan, E. Lai, H. A. Lewis, S. K. Nair and G. A. Petsko for many useful discussions. S.K.B. is an Investigator in the Howard Hughes Medical Institute. This work was supported by the National Institute of General Medical Sciences.

Correspondence and requests for materials should be addressed to S.K.B. (e-mail: burley@rockvax.rockefeller.edu). Atomic coordinates have been deposited to the Protein Data Bank under accession number 1dp7.

Structure of a ligand-binding intermediate in wild-type carbonmonoxy myoglobin

Kelvin Chu[†], Jaroslav Vojtechovsky[‡], Benjamin H. McMahon[§], Robert M. Sweet^{||}, Joel Berendzen[†] & Ilme Schlichting[‡]

^{*}P-21 Biophysics Group, MS-D454, Los Alamos National Laboratory, Los Alamos, New Mexico, 87545, USA

[†]Department of Physics, Cook Building, University of Vermont, Burlington, Vermont 05405, USA.

[‡]Max Planck Institute for Molecular Physiology, Department of Biophysical Chemistry, Otto-Hahn-Str 11, 44227 Dortmund, Germany

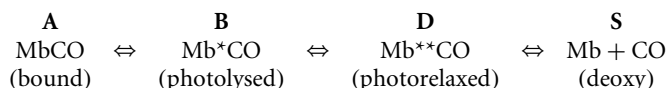
[§]Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, New Mexico, 87545, USA

^{||}Department of Biology, Brookhaven National Laboratory, Upton, New York 11973-5000, USA

Small molecules such as NO, O₂, CO or H₂ are important biological ligands that bind to metalloproteins to function crucially in processes such as signal transduction, respiration and catalysis. A key issue for understanding the regulation of reaction mechanisms in these systems is whether ligands gain access to the binding sites through specific channels and docking sites, or by random diffusion through the protein matrix. A model system for studying this issue is myoglobin, a simple haem protein. Myoglobin has been studied extensively by spectroscopy, crystallography, computation and theory^{1–11}. It serves as an aid to oxygen diffusion but also binds carbon monoxide, a byproduct of endogenous haem catabolism. Molecular dynamics simulations^{3–5}, random mutagenesis⁶ and flash photolysis studies^{7–10} indicate that ligand migration occurs through a limited number of pathways involving docking sites. Here we report the 1.4 Å resolution crystal structure of a ligand-binding intermediate in carbonmonoxy myoglobin that may have far-reaching implications for understanding the dynamics of ligand binding and catalysis.

The ligand-binding reaction in myoglobin (Mb) has been studied by a variety of techniques^{1–14}. Time-resolved and kinetic studies over

wide ranges in time and temperature indicate that ligand binding is not a simple process, but consists of a series of sequential reactions as the ligand moves from the surrounding solvent through the protein to the binding site at the haem iron.



Initially the system has CO bound to the haem iron to form carbonmonoxy myoglobin (A, MbCO). After either thermal- or photodissociation, the CO migrates inside the haem pocket to a primary 'docking site' (B, Mb*CO). At room temperature, mid-infrared bands corresponding to state-B formation are generated within picoseconds after photolysis of MbCO¹¹; as time evolves, the bands disappear either by geminate rebinding (B → A)¹⁰ or by relaxation to a new state. This second intermediate, which has been observed by for sperm whale myoglobin (swMb)¹¹, is also geminate and builds up transiently at room temperature. The secondary docking site can also be populated at low temperature by illumination with intense laser light and is denoted photorelaxed state (D, Mb**CO)¹⁰. The two primary ligand-binding intermediates, states B and D, have median binding enthalpies of 10 and 30 kJ mol⁻¹, respectively¹⁰. At high temperatures and after long times, the ligand escapes the protein matrix to the solvent state (S, Mb + CO, deoxy).

Whether the second, spectroscopically distinct intermediate state is caused by a new docking site or by conformational changes induced in the protein has been the subject of much discussion^{7–14}. A detailed analysis of recombination kinetics reveals that the ligand must be located elsewhere than the distal pocket in state D and suggests that proximal cavities are implicated in the rebinding pathway⁹. To address this issue, we determined the crystal structure of Mb**CO and compared it with that of Mb*CO and MbCO.

We chose horse heart MbCO (hhMbCO) for our study because there is a single ligand-bound conformational substate at neutral pH (ref. 10) whereas sperm whale MbCO (swMbCO) has three taxonomic substates (A₀, A₁, A₃)^{2,11,15} that potentially complicate the analysis. Because of differences in kinetics between the two species, we also determined the structures of hhMbCO and hhMb*CO (Table 1; Fig. 1). In hhMbCO, the ligand is bound almost perpendicular to the haem plane and the haem iron is in plane. In hhMb*CO, the photodissociated CO is located parallel to the haem plane on top of the carbon pyrole at a distance of 3.6 Å from the haem iron. The iron has moved by ~0.25 Å out of the mean haem plane, as described previously for swMb*CO (refs 16, 17). A CO orientation that has the carbon atom closer to the iron was preferred in the refinement. Neither the distal nor the proximal histidine move significantly upon photolysis, as described previously^{17,18}. Spectroscopic measurements of the photolysed state^{11,12} reveal two bands that correspond to CO in different orientations in the distal pocket, one of which converts into the other with a low enthalpic barrier.

In hhMb**CO, the iron is ~0.23 Å out of the mean haem plane (Fig. 2). There is no electron density at the positions of the bound or photolysed CO in the data set collected under illumination (some electron density for bound CO was observed in the data set collected without illumination during data acquisition; otherwise the two complexes are identical), but a new roundish peak appears below the haem plane (see Fig. 1b, c). It was modelled with a CO molecule having 30% occupancy, on the basis of results from equivalent spectroscopic studies carried out in solution. The round electron density assigned to the CO position is in agreement with the observation that the infrared band of state D is broader than that of the state B (refs 10,11,19) and the ligand is probably not oriented within the binding site. The ligand density is sandwiched between the side chains of the proximal His 93 and Phe 138, and surrounded

by Leu 89, Leu 104 and Ile 142. This site has been identified as a binding site for xenon ($\text{Xe}^{(1)}$) (ref. 14) and possibly O_2 (ref. 2). Xenon is an anaesthetic and has been shown to bind preferentially in hydrophobic protein cavities¹⁴. Our data confirm the assignment of the $\text{Xe}^{(1)}$ -binding site as a secondary ligand-docking site, which was suggested from room-temperature flash photolysis studies on the O_2 -rebinding kinetics of wild-type and mutant forms of myoglobin⁸. Two relaxation rates have been observed (half-lives, 20 ns and 0.5–2 μs , respectively), suggesting two distinct transiently

occupied binding sites⁸. The latter is significantly affected in proteins having residues surrounding the $\text{Xe}^{(1)}$ site mutated (for example, Leu 104). Studies probing the formation of state **D** using mutagenesis implicate residues His 97 and Leu 89 (ref. 20). No density is observed in the distal xenon pocket, $\text{Xe}^{(4)}$, as suggested by mutagenesis and computational studies^{4,20,21}.

The trajectory for CO migration from the distal to the proximal side of the haem cannot be inferred from our structures. Anisotropic temperature factors from atomic resolution swMbCO

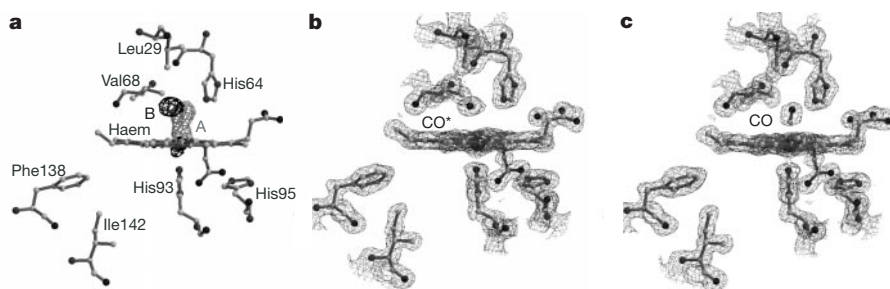


Figure 1 Position of the CO molecule in states **A** and **B**. **a**, Electron-density map of $F_{\text{obs}}(\text{MbCO}^*) - F_{\text{obs}}(\text{CO bound})$ shows positive (black, 10 σ) and negative (grey, 7 σ) electron density at the position of CO^* (**B**) and bound CO (**A**), respectively. **b**, Electron density map of $2F_{\text{obs}} - F_{\text{calc}}$ (1.0 σ) of state **B** generated by photolysis. The iron is located

below the haem plane; the CO molecule is shown almost along its axis. **c**, Electron density map of $2F_{\text{obs}} - F_{\text{calc}}$ (1.0 σ) of state **A**. The iron is in the haem plane; CO is bound almost perpendicular to the haem plane.

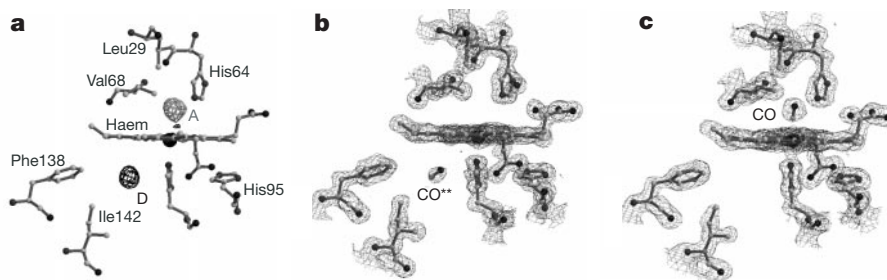


Figure 2 Position of the CO molecule in state **D**. **a**, Electron density map of $F_{\text{obs}}(\text{Mb}^{**}\text{CO}) - F_{\text{obs}}(\text{CO rebound})$ shows positive (black, 4.2 σ) and negative (grey, 5.2 σ) electron density below the haem and at the position of bound CO, respectively. **b**, Electron density maps of $2F_{\text{obs}} - F_{\text{calc}}$ (1.0 σ). The iron is below the haem plane; no electron density associated with

CO is seen at the distal side of the haem, but a new peak interpreted as CO^* appears below the haem. **c**, The CO-bound complex obtained after briefly thawing the photorelaxed complex shown in **a**, **b**.

Table 1 Photolysis protocol, data collection and refinement statistics

	MbCO	Mb*CO	Mb**CO	MbCO (rebound)
Data collection and statistics				
X-ray source, λ	X12C, 1.1	X12C, 0.91	X12C, 0.91	Rotating anode, 1.54
Temperature (K)	100	~22	88	100
Light	None	Fibre, white	Laser, green	None
Specials	None	He-cooling	T-ramp	After unfreezing
$P2_1$ cell: a, b, c ($\text{\AA}$)	63.6, 28.8, 35.6	63.3, 28.8, 35.5	62.8, 28.8, 35.2	63.1, 28.7, 35.3
β ($^\circ$)	$\beta = 106.5$	$\beta = 106.7$	$\beta = 106.2$	$\beta = 106.0$
Resolution ($\text{\AA}$)	22.0–1.4	26.0–1.4	22.0–1.4	27.0–1.7
Completeness (%) [*]	96.1/87.2	94.7/85.9	98.1/98.3	88.8/65.5
Redundancy (%) [*]	2.5/1.6	2.3/1.5	2.4/2.4	2.5/1.4
R_{sym} † (%) [*]	7.0/22.7	7.2/19.5	8.3/26.7	7.3/16.9
$I/\sigma(I)$ [*]	9.2/2.3	7.3/1.6	7.4/2.6	18.7/3.3
Refinement				
Resolution ($\text{\AA}$)	20.0–1.45	20.0–1.45	22.0–1.4	20.0–1.7
$R_{\text{work}}/R_{\text{free}}$ † (%)	21.1/25.5	20.4/25.1	19.7/24.3	19.7/25.9
R.m.s. deviations bond distance ($\text{\AA}$)/angles ($^\circ$)	0.013/1.93	0.007/1.21	0.007/1.23	0.007/1.20
No. of sulphate/water molecules	2/139	2/145	2/163	2/118
CO (occupancy, location)	100%; bound	100%; above haem	~30%; below haem	100%; bound
PDB code	1dwr	1dws	1dwt	

T-ramp, temperature ramp.

^{*} Completeness, redundancy, R_{sym} and $I/\sigma(I)$ are given for all data and for the highest resolution shell: CO-bound, Mb*CO, Mb**CO 1.5–1.4 $\text{\AA}$, CO-rebound 1.8–1.7 $\text{\AA}$.

† $R_{\text{sym}} = \sum |I - \langle I \rangle| / \sum I$.

‡ $R_{\text{work}} = \sum |F_{\text{obs}} - F_{\text{calc}}| / \sum |F_{\text{obs}}|$. The same set of 5% randomly chosen reflections were used for R_{free} calculation.

structures indicate that movement of the haem^{2,22} is primarily in-plane, suggesting that a transient opening to the proximal cavity due to a sliding haem with concomitant motions of adjacent side chains may lie along the reaction path. This would be consistent with the proposal that relaxation is most probably triggered by the relaxation of the haem^{10,23,24}. In general, computation studies show the ligand hopping between the four xenon-binding cavities before escaping to the solvent. Indeed, simulations of CO diffusion through Mb^{3,5,8} show the possibility of trajectories leading from the distal pocket to the Xe⁽¹⁾ site.

Why should such a simple reaction as oxygen binding require such highly engineered and evolutionarily conserved docking sites? The answer may lie in the fact that dissolved O₂ must compete with water to bind to Mb. Deoxy Mb has a water molecule inside the distal pocket near the state B site. As the concentration of water is 10⁵ times that of O₂ in oxygen-saturated solutions, state D may serve as a local 'storehouse' of O₂ near the active site, thereby increasing the effective O₂ concentration many times. Thus, cavities or packing defects in Mb appear to have a functional role in ligand binding and may be important not only for the kinetics of ligand binding but also for determining relative affinities. As the fluctuations connecting the cavities determine accessibility, picosecond fluctuations of side chains may have an evolutionarily conserved survival value²⁵. The role of hydrophobic cavities as a means of controlling protein function may be a general feature in ligand binding in proteins as suggested earlier for Ni-Fe hydrogenase²⁶. □

Methods

Horse heart Mb (Sigma) was crystallized at room temperature by equilibrating 10 µl drops of 5 mg ml⁻¹ protein in 1.7–1.8 M ammonium sulphate and 0.1 M Tris-HCl pH 7.5 against 1 ml of 3.4–3.6 M ammonium sulphate and 0.1 M Tris-HCl pH 7.4 using hanging-drop geometry. Rosette-shaped crystals grew within 2 weeks and leaflets (typically 0.01 × 0.07 × 0.3 mm) were separated for data collection. The hhMbCO crystals were prepared by soaking the leaflets in CO-saturated mother liquor supplemented with 8 mg ml⁻¹ dithionite, 5 µl 2 M NaOH ml⁻¹ and 7.5 % glycerol until a colour change occurred from brown to raspberry red. Subsequently, the crystals were flash cooled in liquid nitrogen. For data collection of the photorelaxed hhMb**CO complex a hhMbCO crystal was illuminated from both sides of the flat surface of the leaflet with an Ar⁺-ion laser for 6 h at 160 K (intensity 4.6 mW mm⁻²), and then the temperature was ramped to 88 K at 10 K h⁻¹ under continuous illumination. The diffraction data were collected under continuous illumination at beamline X12C at the NSLS using the 2k × 2k Brandeis CCD detector. The laser was then switched off and another dataset was collected (data not shown). After collection of the two hhMb**CO data sets, the crystal was unfrozen for ~30 s by blocking the Oxford cryostream before a third dataset was collected of the same crystal using a rotating anode and a Siemens Hi-Star detector (hhMbCO rebound). The data of the photolysed hhMb*CO complex were collected with the 2k × 2k Brandeis CCD detector at ~22 K using an open flow helium cryostat to make rebinding negligible during data collection. Before data collection, the crystal was illuminated for 30 min, and then diffraction data were collected under continuous illumination with white light obtained from a fibre illuminator as described¹⁶.

The diffraction data were reduced with XDS²⁷ and scaled with XSCALE. The structures were refined with XPLOR3.851 (ref. 28) using PDB entry 1AZI as a starting model omitting water molecules, the ligand and sulphate ions. Refinement included simulated annealing to 2,200 K and individual B-factor refinement. Model building was done with the program O²⁹. The CO molecules were inserted at the last step of the refinement. Figures were prepared with Bobscript³⁰.

Received 3 November; accepted 16 December 1999.

1. Antonini, E. and Brunori, M. *Hemoglobin and Myoglobin in their Reactions with Ligands* (North-Holland Pub. Co., Amsterdam, 1971).
2. Vojtechovsky J. *et al.* Crystal structures of myoglobin-ligand complexes at near-atomic resolution. *Biochem. J.* **77**, 2153–2174 (1999).

3. Elber R. and Karplus, M. Multiple-conformational states of proteins: a molecular dynamics analysis of myoglobin. *Science*, **235**, 318–321 (1987).
4. Carlson, M. L., Regan, R. M. & Gibson, Q. H. Distal cavity fluctuations in myoglobin: protein motion and ligand diffusion. *Biochem. J.* **35**, 1125–1136 (1996).
5. Tilton, R. J. Jr *et al.* Computational studies of the interaction of myoglobin and xenon. *J. Mol. Biol.* **192**, 443–456 (1986).
6. Huang X. & Boxer S. G. Discovery of new ligand-binding pathways in myoglobin by random mutagenesis. *Nature Struct. Biol.* **1**, 226–229 (1994).
7. Austin R. H. *et al.* Dynamics of ligand-binding to myoglobin. *Biochem. J.* **14**, 5355–5373 (1975).
8. Scott, E. E. & Gibson Q. H. Ligand migration in sperm whale myoglobin. *Biochemistry* **36**, 11909–11917 (1997).
9. McMahon, B. H. Energetics of protein fluctuations: Ligand binding to myoglobin and electron transfer in reaction center. Thesis, Univ. of Illinois at Urbana-Champaign, (1997).
10. Chu, K. *et al.* Light-induced and thermal relaxation in a protein. *Phys. Rev. Lett.* **74**, 2607–2610 (1995).
11. Lim, M., Jackson, T. A. & Anfinsen, P. A. Ultrafast rotation and trapping of carbon monoxide dissociated from myoglobin. *Nature Struct. Biol.* **4**, 209–214 (1997).
12. Alben, J. O. *et al.* Infrared spectroscopy of photodissociated myoglobin. *Proc. Natl Acad. Sci. USA* **79**, 3744–3748 (1982).
13. Powers, L. *et al.* Kinetic, structural and spectroscopic identification of geminate states of myoglobin: a ligand-binding site on the reaction pathway. *Biochemistry* **26**, 4785–4796 (1987).
14. Tilton, R. F. Jr., Kuntz, I. D. Jr. & Petsko, G. A. Cavities in proteins: structure of a metmyoglobin-xenon complex solved to 1.9 Å. *Biochemistry* **23**, 2849–2857 (1984).
15. Frauenfelder, H., Sligar, S. G. & Wolynes, P. G. The energy landscapes and motions of proteins. *Science* **254**, 1598–1603 (1991).
16. Schlichting, L., Berendzen, J., Phillips, G. N. Jr. & Sweet, R. M. Crystal structure of photolysed myoglobin. *Nature*, **371**, 808–812 (1994).
17. Hartmann, H. *et al.* X-ray structure determination of a metastable state of carbonmonoxymyoglobin after photodissociation. *Proc. Natl Acad. Sci. USA*, **93**, 7013–7016 (1996).
18. Teng, T. Y., Srajer, V. & Moffat, K. Photolysis-induced structural changes in single crystals of carbonmonoxymyoglobin at 40 K. *Nature Struct. Biol.* **1**, 701–705 (1994).
19. Anfinsen, P., deVivie-Riedle, R. & Engel, V. Ultrafast control and detection of molecular dynamics. *Proc. Natl Acad. Sci. USA* **96**, 8328–8329 (1999).
20. Gibson, Q. H., Regan, R., Elber, R., Olson, J. S. & Carver, T. E. Distal pocket residues affect picosecond ligand recombination in myoglobin: an experimental and molecular dynamics study of position 29 mutants. *J. Biol. Chem.* **267**, 22022–22034 (1992).
21. Quillin, M. L. *et al.* Structural and functional effects of the distal valine in myoglobin. *J. Mol. Biol.* **245**, 416–436 (1995).
22. Kachalova, G. S., Popov, A. N. & Bartunik, H. D. A steric mechanism for inhibition of CO binding to haem proteins. *Science* **284**, 473–476 (1999).
23. Abadan, Y. *et al.* Ligand binding to haem proteins V: light-induced relaxation in proximal mutants L89I and H97F of carbonmonoxymyoglobin. *Biophys. J.* **68**, 2497–2504 (1995).
24. Ahmed, A. M. *et al.* Evidence for proximal control of ligand specificity in haemproteins: absorption and Raman studies of cryogenically trapped photoproducts of ligand bound myoglobins. *Chem. Phys.* **158**, 329–352 (1991).
25. Brunori, M. *et al.* Does picosecond protein dynamics have survival value? *Trends Biochem. Sci.* **24**, 253–255 (1999).
26. Montet, Y. *et al.* Gas access to the active site of Ni-Fe hydrogenases probed by X-ray crystallography and molecular dynamics. *Nature Struct. Biol.* **4**, 523–526 (1997).
27. Kabsch, W. Automatic processing of rotation diffraction data from crystals of originally unknown symmetry and cell constants. *J. Appl. Crystallogr.* **24**, 795–800 (1993).
28. Brünger, A. T. *X-PLOR: A System for Crystallography and NMR*, Version 3.1 (Yale Univ. Press, New Haven, 1992).
29. Jones, T. A. *et al.* Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).
30. Esnouf, R. A. An extensively modified version of MolScript that includes greatly enhanced coloring capabilities. *J. Mol. Graphics* **15**, 132–134 (1997).

Acknowledgements

The work was generously supported by the Human Frontiers Science Program (to I.S. and J.B.), the Bundesministerium für Bildung und Forschung (to I.S.) and the Richard and Anne-Liese Gielen-Leyendecker-Stiftung (to I.S.). Beamline X12C is supported by the US Department of Energy Offices of Health and Environmental Research and of Basic Energy Sciences, and by the National Science Foundation. We are grateful to H. Frauenfelder for stimulating discussions, G. Holtermann for expert technical assistance, and R. S. Goody for continuous support and encouragement.

Correspondence and requests for materials should be addressed to K.C. (e-mail: kelvin.chu@uvm.edu) or I.S. (e-mail: ilme.schlichting@mpi-dortmund.mpg.de). The coordinates have been deposited in the PDB under accession numbers 1dwr, 1dws, 1dwt.

The heat of the moment

Thermocyclers, and other manipulators and recorders of temperature.

Touchgene

Technew www.technew.co.uk & www.technusa.com

A thermal cycler designed for ease of use and flexibility

Touchgene incorporates a 115 × 90 touch-sensitive screen providing a user-friendly programming interface showing a real-time plot of both block and calculated sample temperature. A wide selection of block options up to a 384-well microplate is available. Up to 100 programs can be stored on each card and each one can be named and password protected. Throughput of samples is quick, Technew say a 30-cycle programme can be completed in well under an hour. Touchgene comes complete with a new universal quick-release heated lid that allows the user to regulate the exact pressure required dependant on the block format being used. This lid also incorporates an internal anti-condensation skirt so that the smallest sample volumes may be used without risk of sample loss onto the lid and sidewalls of the sample tube.

Reader Service No. 100

Gene Cycler

Bio-Rad

www.bio-rad.com

A thermal cycler that gets personal

The Gene Cycler unit is an inexpensive personal mini thermal cycler, fully licensed for the PCR process. The machine has a footprint small enough for it to fit on a researcher's desk. It's an advanced solid-state device using electrical resistors for heating and a fan for cooling, resulting in fast ramping times. The temperature displayed corresponds to the temperature inside the tube, because the cycler uses software algorithms to compensate for temperature differences between the heating block and the sample. The cycler has the ability to link up to 6 programs and still provides enough memory to store up to 100 programs.

Reader Service No. 101

TC 1600

Intelligent Bio-Instruments

www.intelligentbio.com

Sixteen-block parallel thermal cycling

The TC 1600 is a high-throughput thermocycler for assays requiring simultaneous heating and cooling of up to 6,144 samples. Using 16 blocks for standard 96/192/384-well microplates, the thermocycler runs four sets of independent heating and cooling protocols. It stores and recalls predefined cycling protocols and allows up to 50 set point temperatures from 30°C to 100°C with looping and pausing.

Reader Service No. 102



Screen test: Touchgene can be sensitive.

Precertified thermometer

Brooklyn Thermometer Company

www.brooklynthermometer.com

A hardy-but-sensitive thermometer

This thermometer, Catalog no. 83406, comes pre-certified with NIST traceability. Usually the customer chooses a thermometer and specifies temperature points for Factory Certification, but this one is pre-certified traceable to NIST at 0, 37, 56 and 100°C. Observed corrections are expressed to the nearest 0.1°C. The 12-inch-long thermometer is ranged -10 to 110°C in 1° divisions and is tested for 76 mm immersion. The '83406' is engraved using Brooklyn's Permanent Markings process, which fuses the graduated lines and numbers directly to the glass making them easier to read than etched thermometers. PM markings are not washed out by exposure to caustic liquids.

Reader Service No. 103

Whatman Biometra

<http://biometra.de/>

Built for speed and uniform temperature

The T1 thermocycler boasts speed and temperature uniformity achieved by combining the silver block concept with the latest in Peltier architecture. Programming is via a spreadsheet-like programming window. Navigation is simple using four cursor keys, and up to 1,000 protocols can be stored in 10 individual subdirectories. The T1 Thermocycler can be controlled from a PC — up to 8 instruments can be controlled from the one computer. Temperature profiles can be recorded and archived for GLP conforming documentation

Reader Service No. 106

Heating elements

MHI Inc.

www.mhi-inc.com

Warm to the task with shapely heating elements

Micropyretics Heaters International Inc. of Cincinnati, have released a new set of complex shaped electrical heating elements for different applications. These include heated pipes for glass blowing and specialized heaters for superconductor and vapour processing. The heaters function to up to 1,900°C in air and may be tailored to have low to high resistance. The heaters may be bought separately or with complete control electronics.

Reader Service No. 104

Ovens and incubators

Lenton Furnaces

www.lentonfurnaces.com

Bench-top ovens and incubators good to $\pm 0.2^\circ\text{C}$

Based in Sheffield, a city that used to be permanently lit by the furnaces of the steel industry, Lenton Furnaces have added a new line of laboratory-scale ovens and incubators to their range of thermal equipment. They meet UK and European standards for health, safety and electro-magnetic emissions. On most models temperature stability, with PID control, is $\pm 0.2^\circ\text{C}$, and it is no more than $\pm 0.5^\circ\text{C}$ on any model. Polished stainless steel interiors are standard, and high-grade insulation is used to reduce power requirements. Incubators and ovens are both available in four bench-top sizes from 27 to 215 litres. They can be supplied with or without fans for air circulation. Options available include specialized versions to cater for volatile or wet materials, and to allow probes and gases to be used.

Reader Service No. 105



One of eight: PC control for the T1 Thermocycler.

new on the market

Ovens and incubators

Carbolite www.carbolite.co.uk

Ovens and incubators, now with process timer

Carbolite, like the company in the preceding item, make ovens and incubators close to Sheffield, in the Hope Valley. The Peak series of general purpose ovens is offered with a wide variety of chamber capacities. Four bench mounted units are available with a choice of gravity or forced air circulation with a maximum operating temperature of 300°C. The two larger floor standing units feature forced air circulation only and offer a maximum operating temperature of 250°C. The Peak range of incubators is available in the same capacities as the ovens, and has a maximum operating temperature of 80°C. A glass inspection door is fitted to allow viewing of the samples without disturbance. All models are designed for good temperature uniformity, and rapid heat-up and recovery. Controllers on Carbolite ovens and incubators now incorporate a process timer function to allow heating and cooling periods to be pre-set or heating to be delayed until a pre-set time.

Reader Service No. 107

EchoTherm

Torrey Pines Scientific
www.torrey-pines.com

Hot and cold running experiments on the lab bench

EchoTherm Model IC20 Electronic Chilling/Heating Plate is a compact machine designed for personal, bench-top use for chilling, freezing, or heating biological and other samples. The unit accepts 96-well assay plates, and is fitted with aluminium blocks made to accommodate 0.2 ml, 0.5 ml and 1.5 ml centrifuge tubes. Potential applications include maintaining 14°C for ligation reactions, storing samples at ice bucket temperatures without the need for ice, storing enzymes or DNA

libraries at the workstation, and heating or chilling a microscope stage. The IC20 is Peltier driven, controls temperature to $\pm 1^\circ\text{C}$, and has a two line alphanumeric display that shows target and actual plate temperatures.

Reader Service No. 108

Mastercycler

Eppendorf www.eppendorf.com

A large-capacity thermal cycler

This model has been introduced as a 'big brother' to the 'Mastercycler personal'. It features a universal block in the form of a 96-well grid and can accommodate 96 0.2-ml tubes, 77 0.5 ml tubes or one microtitre plate. Method transfer between different Mastercycler workstations is simplified by the use of credit-card style memory cards, so a protocol does not need to be re-entered. Programming includes adjustable ramps as well as time and temperature ramps, and the Autorestart option guards against problems caused by power failure.

Reader Service No. 109

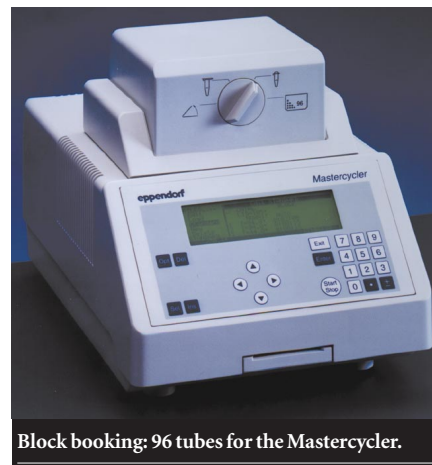
Column heater/cooler

BAS www.bioanalytical.com

Thermostatic control over liquid chromatography columns

The Igloo-CIL column temperature controller can be used at either elevated or reduced temperatures. It uses the Peltier effect to maximize temperature stability and reproducibility. An automated calibration system controls the temperature display and actual internal temperature. A constant temperature in HPLC is important to minimize detector drift and optimize reproducibility. The Igloo-CIL column thermostat can hold up to six analytical columns or three semipreparative (such as GPC) columns at one time and offers a temperature range of 0–99°C, in 0.1 °C increments, including true ambient.

Reader Service No. 110



Block booking: 96 tubes for the Mastercycler.

Thermocouple/Thermowell

Eagle Stainless Containers
www.eaglestainless.com

Monitor temperatures inside a container

This new Thermocouple/Thermowell is built to monitor and record temperatures inside a container, and the temperature of its contents, for extended periods of time. The device can be used for tracking temperatures during shipping, storage or process. The Thermowell is fabricated from T316L stainless steel and has a surface finish of 24Ra. The Thermocouple, when used in conjunction with the TempTale 3, mounts securely inside the Thermowell and provides an accurate reading and recording of the interior temperature of a container/bottle and its contents. The data can then be downloaded to a PC to display the temperatures over the time period. With the TempTale and Thermocouple removed the bottle/container can be sterilized in an autoclave.

Reader Service No. 111

HE10D, HE30D

Grant Instruments www.grant.co.uk

A nice warm bath?

These two high-temperature laboratory baths and circulators, of 10 and 30 litre capacity, are designed for use between 50°C and 260°C. In the HED series, localized overheating, sometimes a hazard with the silicone oil immersion fluid commonly used for high temperatures, is minimized by the use of a very powerful stirrer and a large surface area heater. Temperature uniformity across the bath is $\pm 0.2^\circ\text{C}$ at 150°C. For safety, these baths are heavily insulated — including the lid — and other safety features include an adjustable temperature cut out and fluid level sensor.

Reader Service No. 112

These notes are compiled in the Nature office from information provided by the manufacturers. For more details, fill in the reader service card, bound elsewhere in the journal.

ADVERTISEMENTS

Check out our homepage at <http://www.resgen.com>

BACs
YACs
PACs
cDNAs

Research Genetics provides clones and colony membranes for a number of human, mouse, and rat libraries as well as a custom screening service for the libraries.

Research Genetics, Inc.
U.S. or Canada 800-533-4363
U.K. 0-800-89-1393
FAX 256-536-9016

READER ENQUIRY NO. 4

Biomolecular Synthesis
oligos, synthetic
peptides and polyclonal
antibodies

Contact: Severn Biotech Ltd
Tel: 01562 825 288
Fax: 01562 825 284
email: info@severnbiotech.com
www.severnbiotech.com

READER ENQUIRY NO. 28